Journal Subline
LNBI 3939

# Transactions on Computational Systems Biology IV

Corrado Priami
Editor-in-Chief

Springer

Lecture Notes in Bioinformatics 3939

Edited by S. Istrail, P. Pevzner, and M. Waterman



Subseries of Lecture Notes in Computer Science

Corrado Priami Luca Cardelli
Stephen Emmott (Eds.)

# Transactions on Computational Systems Biology IV

Springer

Series Editors

Sorin Istrail, Brown University, Providence, RI, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Corrado Priami
The Microsoft Research - University of Trento
Centre for Computational and Systems Biology
Piazza Manci, 17, 38050 Povo (TN), Italy
E-mail: priami@msr-unitn.unitn.it

Luca Cardelli
Stephen Emmott
Microsoft Research Cambridge
7 JJ Thomson Avenue, Cambridge CB3 0FB, UK
E-mail: {luca,semmott}@microsoft.com

Library of Congress Control Number: 2006923001

CR Subject Classification (1998): J.3, H.2.8, F.1

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-33245-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-33245-9 Springer Berlin Heidelberg New York



Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11732488 06/3142 5 4 3 2 1 0

# Preface

This issue of the journal reports some selected contributions from the First Converging Science conference held in Trento in December 2004 and chaired by Luca Cardelli, Stephen Emmott and Corrado Priami. The first contribution is the transcription of the keynote lecture by Robin Milner on the foundations of global computing. The second paper, by John Wooley, reports on the interdisciplinarity in innovation initiatives. Ronan Sleep presents a grand challenge for the convergence of sciences in the third paper, while Andrew Jones discusses how to apply computer science research to biodiversity in the fourth contribution. The fifth paper, by Bob Hertzberger, introduces the concept of e-science. The sixth paper, by Francois Fages, describes the role of syntax and semantics of programming languages in systems biology, while Ivan Arisi and his co-authors report on a European initiative on neuronal cells in the seventh contribution. The eighth paper, by Imrich Chlamtac et al., deals with the role of biological inspiration in autonomic computing. The ninth paper, by Riguidel, covers digitalization and telecommunications.

The volume ends with two regular contributions. The first one by Blossey, Cardelli and Phillips describes a compositional approach to the quantitative modelling of gene networks. The last paper of the volume by Jeong and Miyano introduces a prediction method for the protein–RNA interaction.

January 2006 Corrado Priami

# LNCS Transactions on Computational Systems Biology – Editorial Board

# Table of Contents

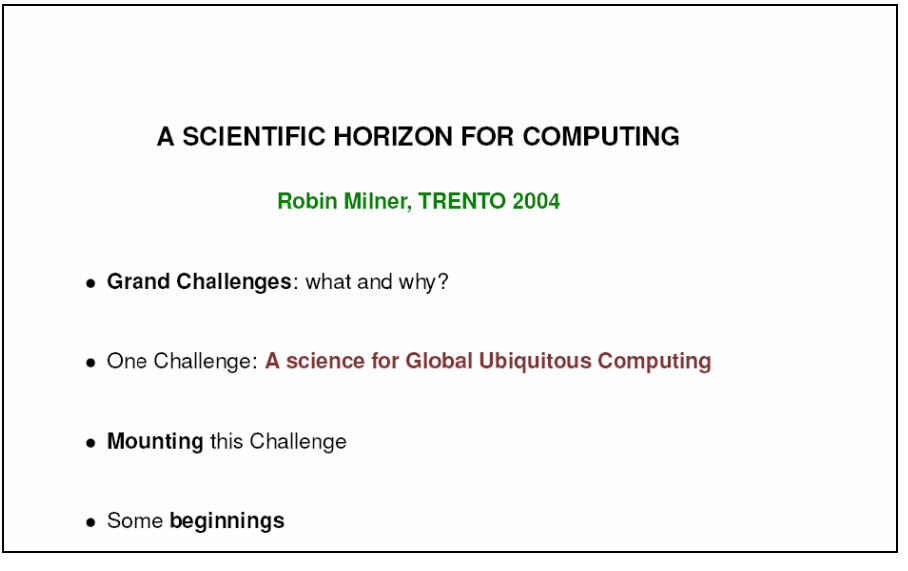
# Transcription of the Presentation
# Scientific Foundation for Global Computing

Robin Milner

University of Cambridge, UK

It is a big honour to be able to speak at one of the most exciting conferences I have been to for 20 years.

I feel in some ways daunted because, although the development of connections between biology and computer science may seem wonderful from outside, we know that they depend crucially on details. We have to believe with confidence that what we have already done justifies bringing the subjects together in this way. I think that's we are going to see in the later talks. I want to try to anticipate a little bit of that here. But first I want to talk about global computing. We may describe it, fancifully perhaps, as the arrival of a single global computer, which is increasingly pervading lives.

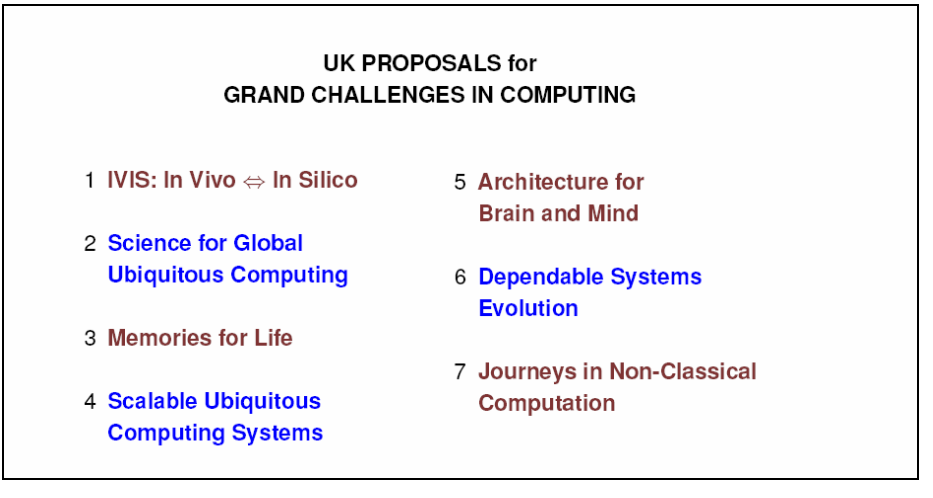



The sequence in my talk will be this: first I am going to put the global computer in the context of the UK exercise on Grand Challenges for Research in Computing, to create a framework for it. Secondly, I am going to focus on a specific Challenge: to build a science that will underlie all this pervasive computing. (We would like to trust it; but are we ready to trust computing systems on such a large scale?) Thirdly: How shall we put more structure into developing this science? What ingredients do we

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 1 – 13, 2006.

already have? And fourthly, I am going to talk about some beginnings for this science; this gives me a sense of security, as it indicates that we are starting from somewhere.

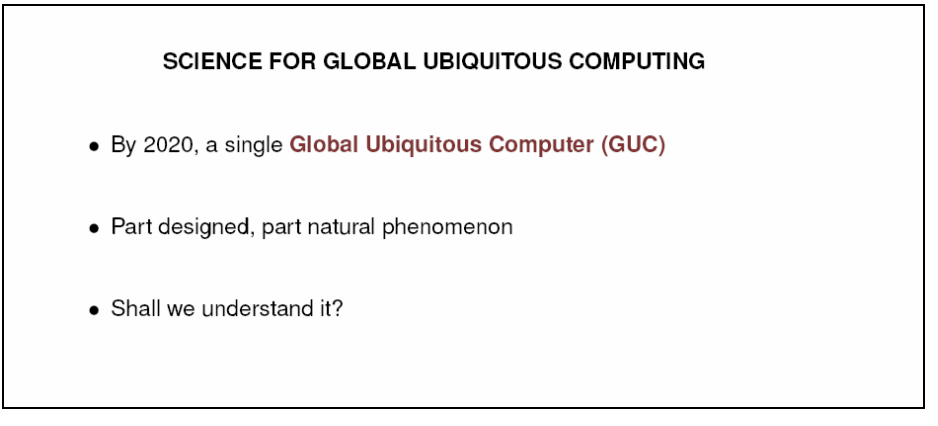

What is a Grand Challenges Exercise? We embark upon it because we believe that long-term challenges for computer science can't just be written down on a piece of paper in five minutes; you have to work towards them. You have to bring together some ingredients into a goal which is so well-focussed that you can devise a plan to reach that goal over 15 years. This focussing takes time, perhaps five years, and it has to come from the community - it can't be dictated from above. The community examines a portfolio of *proposals* for long-term challenges; some of them are successfully developed into Grand Challenges – something of which we can say "Yes! We can *plan* to achieve this in 15 years." Of course it still has a reasonable possibility of failure, indeed, the plan must define what would count as failure.



Such an exercise helps us to identify, within our subject community, what our real aims are; it also has the benefit of showing the public and the funding agencies what we aspire to. We must cease to think just in terms of building the next technology. We must have aspirations, defining the directions we wish to take.

So we have developed a portfolio of Grand Challenge proposals. Rather conveniently, two among them are perfectly relevant to what we shall be discussing in these two days. The first, *In Vivo--In Silico*, has to do with bringing computing into biology or vice versa. It is co-ordinated by Ronan Sleep, who is here with us; he has asked me to say that there are some handouts to do with this Grand Challenge at the desk outside. So that is the first Challenge on our list; it wouldn't exist if it were not for the success of some small but very indicative predictive experiments, bringing computing models into biology.

The second one, *Science for Global Ubiquitous Computing*, is the one I want to start from. I want to start there because I know something about how to predict what should happen there, but also because one of the things we would like to see is a convergence between the models used on the one hand to *engineer* pervasive computing, and those used on the other hand to *understand* existing scientific -- possibly biological -- systems. It's almost too much to hope that the principles for those two things are the same. But let us at least entertain that hypothesis, because this is one of the most exciting things that could happen. We see such disparate fields of study as possibly having the same underlying very particular primitive ideas. So I want to have time to say something about that later in my talk.




First then, what about the *Science for Global Ubiquitous Computing*? You could imagine we have a single computer, or network of computing entities, that pervades the world by 2020. Who's to say that it will not happen? It is going to be partly designed and partly a natural phenomenon, because it comes together from the energies of different people collaborating or possibly just simply doing their own thing and then joining their systems up. The question is: Shall we understand it? We have problems understanding our own software. For example, recently in UK we have the glorious failure of the government's Child Support Agency to produce a decent computing system. Grand failures like this are common.

**UNDERSTANDING and BUILDING**

- Underlying both are **modelling kits**
- **Traditional science and engineering** has
  *Differential equations, Laplace transforms, Matrix algebra, . . .*
  . . . and they join understanding with building
- **Computer science and engineering** has
  *Automata, Languages, Relational algebra, Network theories, Logics, Stochastics, Type theory, Process calculi, Semi-structured data, . . .*
  . . . but the junction is tenuous. **Why?**
- For **Ubiquity??** Separation will lead to *stagnation or worse.*

So the key, for me, is that understanding and building have to go together. Both science and engineering involve modelling kits. In standard mathematics and engineering it is easy to reel off a list of these things: differential equations, Laplace transforms etc. Computer science also has a number of theoretical models, which could claim this kind of status; but the fact is that those theories are not really used in the engineering of computing systems. Why not? Well, the theories struggle along trying to keep up with the latest technology. I think that's why it is difficult for the theories to inform engineering practice. Our technologies, both hardware and software, move so fast and demand new ideas in such a way that new theories have to emerge to try to keep up with them. So, even now, the junction is tenuous between the science and the engineering of computing systems; this situation will be amplified as we move forward into global computing. But the continued separation between science and engineering will lead to stagnation of systems, or even to disaster.

**The Challenge: SCIENCE FOR GLOBAL UBIQUITOUS COMPUTING**

- *To develop an informatic science whose* **concepts**, **calculi**, **theories** *and automated* **tools** *allow descriptive and predictive analysis of the GUC at each level of* **abstraction**
- *That every* **system** *and* **software** *construction —including languages— for the GUC shall employ only these concepts and calculi, and be* **analysed** *and* **justified** *by these theories and tools.*

www.nesc.ac.uk/esi/events/
Grand_Challenges/proposals/Ubiq.pdf

An *ideal* goal? But no argument limits the degree of possible success!

In our proposal we have boldly formulated a couple of statements which should be the Grand Challenge for global computing:

- *To develop an informatic science whose concepts, calculi, theories and automatic tools allow descriptive and predictive analysis of the global ubiquitous computer at each level of abstraction.*
- *That every system and every software construction, including languages for global computing, shall employ only those concepts and calculi, and shall be analysed and justified by those theories and tools.*

That is the goal that we want realise. It is an ideal goal, but there is no argument that limits the degree of possible success in that direction. That is the important point: not that we won't fail, but there is no present reason that you could claim *why* we should fail, except for dragging our feet.

**A THEORETICAL HIERARCHY**

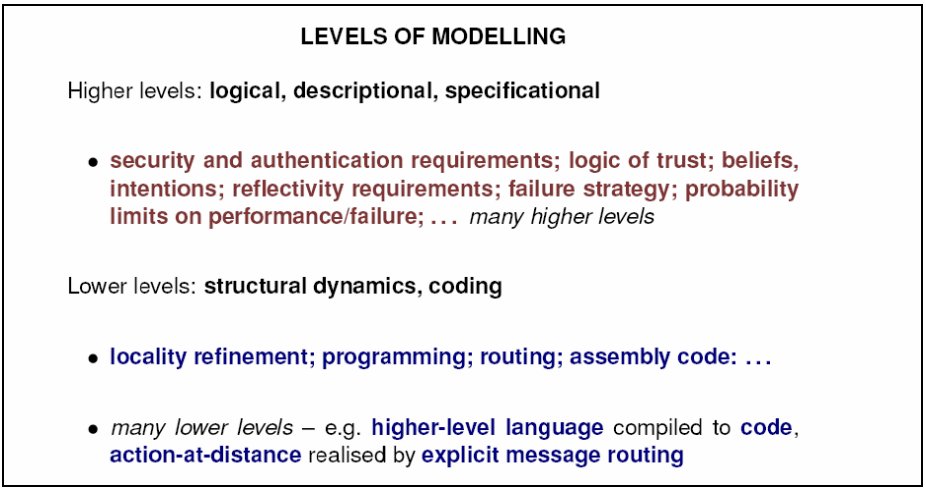
**Theoretical** goals for the Grand Challenge:

- *To express theories for the GUC as a hierarchy of* **models** *and* **languages**, *assigning each concept (*e.g. *trust) to a certain level in the hierarchy*
- *To define, for each model $M$, how a system description in $M$ may be* **realised** *or* **implemented** *in models/languages $M_1, \ldots, M_n$ lying below $M$*

Why do we need models at many abstraction levels?

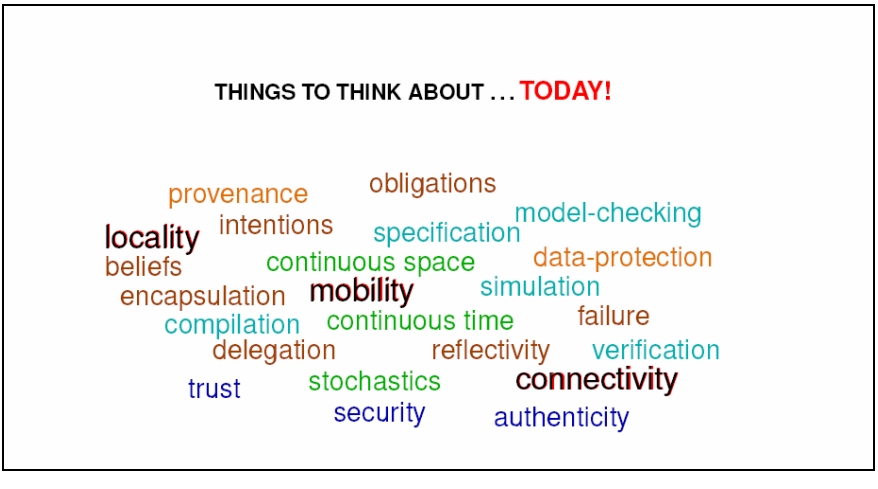
The theory behind this will have to build a hierarchy of models and languages; it will somehow assign each concept, such as *trust* or *failure* or *interactivity* or *logical analysis* or *software development*, to a certain level in the hierarchy. Then, at each level of modelling, it must define how a system described at that level can be realised or implemented in terms of languages or models which lie below it. That is of course a very clean picture; it is never going to look quite like that. But a hierarchy of models is going to be essential, because we shall be dealing with a huge population -- perhaps billions – of computing entities.

We need many abstraction levels. I am not going to go into great detail about them, but they must follow a pattern. First of all, at a higher level we tend to be descriptional; we tend to specify how systems should behave; we tend to analyse them logically, often in terms of rather high level entities like trust or in terms of certain invariant properties that represent acceptable behaviour. At intermediate levels there might be strategies for failure management, probability limits on performance of failure, reflectivity requirements (the ability for a system to report on what it has



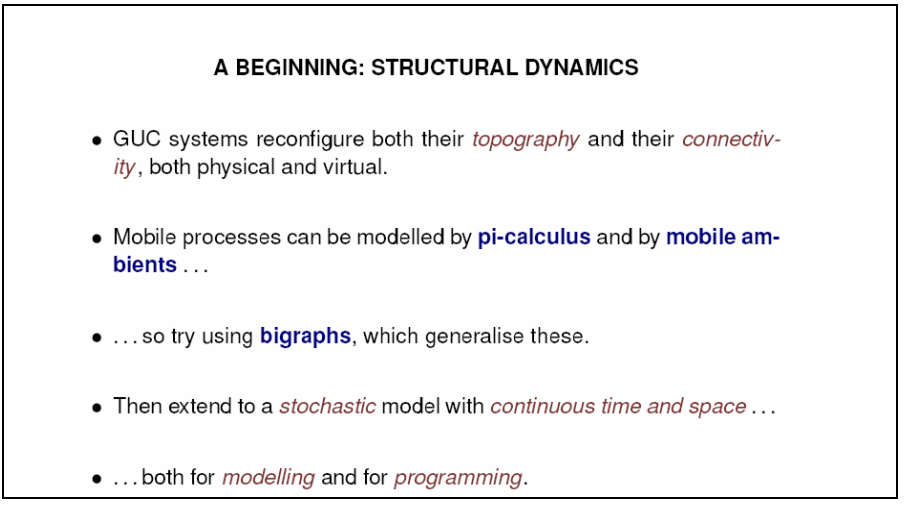

done). At the lower levels we have structural dynamics, how things actually work in practice -- including machine code, routing of messages and so on. So you can see the kind of things that this levelling consists of. I don't believe that we can build a theory without such levels. I even believe that there should be a fractal quality here; for example, movement of data within a computer program should be treated in the same way as movement of agents in a natural or built environment. There should be a certain repetition of concepts among the levels; how else can we understand a huge organism with a manageable repertoire of concepts?

I think that's enough about levels of modelling.




This slide shows just some of the concepts we have to consider, things like data protection, or the intentions of a individual human or software agent, or delegation of work from one system to another. We also have to bring in an understanding of continuous time and space, and our analysis must have a probabilistic or stochastic

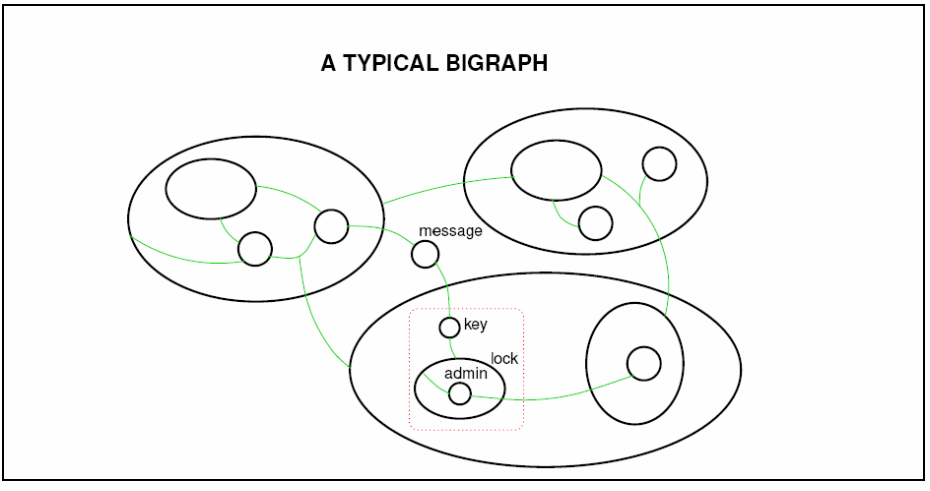
element. You see this huge space of things. I want to pick out three things that seem to lie where our collaboration with biology begins, and also where we begin to understand global computing. They are: *locality*, *connectivity*, and *mobility*.



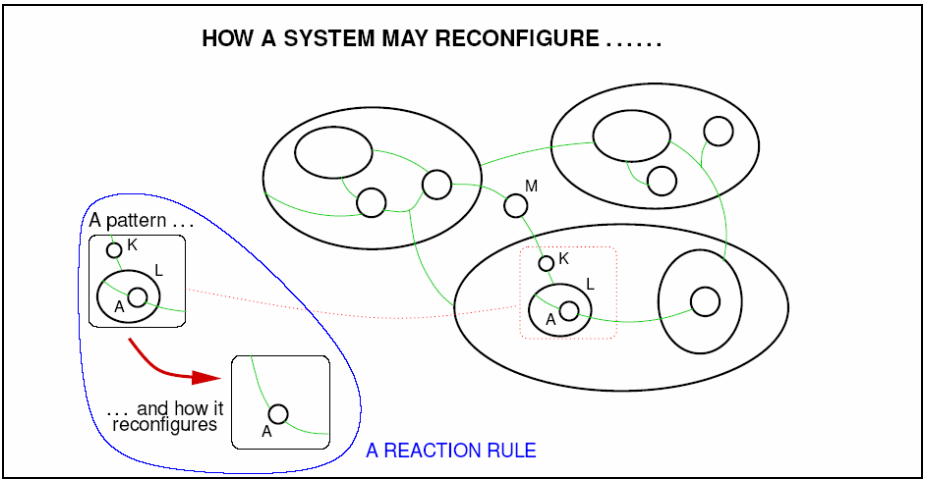

*Locality*, or topography, is of course an essential feature of computing systems that may be embedded in our environment or even in our bodies. It may appear to be a new concern in computing, but in fact we have always used topographical intuitions in our computer programs. For example, we talk of storage *space*, or *local* variables. It even looks as though the macroscopic or real-life topographic considerations that come with global computing may line up with the microscopic or virtual topography that lives inside a computer program. That would be a wonderful junction. But there is more to this space than just locations. By programming, we achieve a situation in which entities located far apart may nevertheless be connected; their means of connection may be ultimately a matter of physical fibres or wireless, but at an abstract level we think of these distant neighbours as if they were next-door neighbours; that is how we understand accessing remotely located websites, for example. Thus *connectivity* can be usefully seen as independent of locality.

To complete the trio, a system may freely reconfigure both its localities and its connectivity; that is what we mean by *mobility*. Do we get the same kinds of mobility in physical and in virtual space? Do we get the same kinds both in artificial (computing) and in natural (biological) systems? If so, we would have a new and exciting theory of dynamical systems, with emphasis upon discrete entities. Moreover, we would hope to apply these ideas equally at both high and low (abstract and concrete) levels.

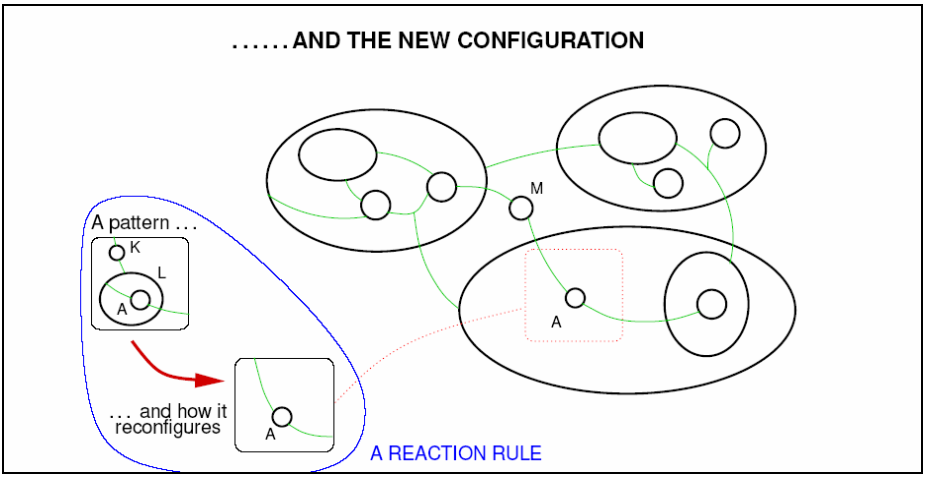
Recently, two calculi of discrete dynamics from computer science have been applied with some success to both biology and pervasive computing; these are the Pi calculus and the calculus of Mobile Ambients. Roughly speaking, they deal respectively with connectivity and with locality. So I am now working with a model called Bigraphs, which tries to capture the best parts of those. The prefix “bi-” refers to the two structural elements: *placing* (locality) and *linking* (connectivity).



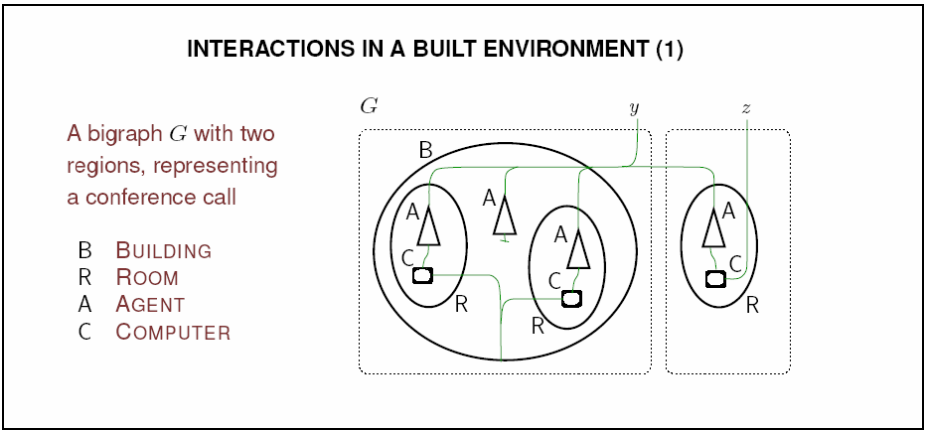

Here is what I call a bigraph. This is a system in which you have things nested inside each other, but they may also communicate across those boundaries irrespective of where they reside. So these green links in the picture connect anything to anything else. I have called some of the nodes *message*, *key* and *lock*, because I wanted to indicate that this might be a system in which a message is trying to insert a key into a lock, in order to find its way to a receiver in the bottom right-hand corner.



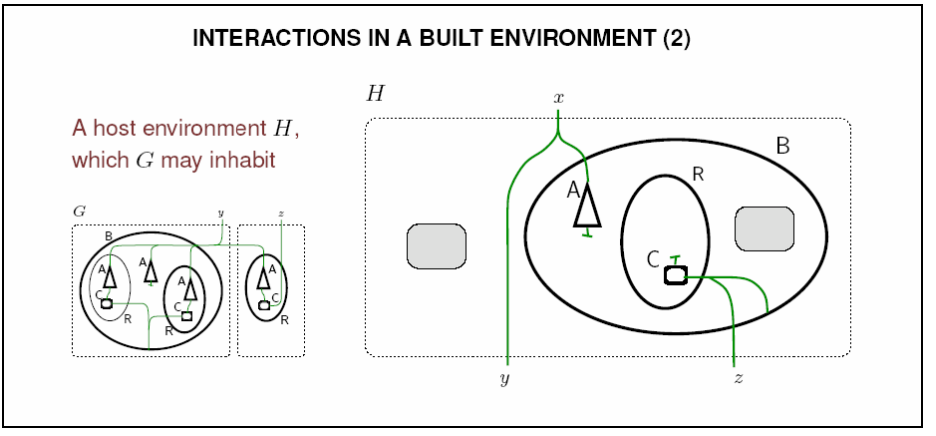

In more detail, here is the message **M**, right in the middle of the picture, the key **K** and the lock **L**; and in the bottom left-hand box is what I call a *reaction rule*. Think of it as a very elementary piece of dynamics; it says that whenever you have a pattern of **K** next to **L**, with an agent **A** inside there, then the pattern can change, the lock and key vanish and you are left with the agent **A** ready to help the message **M** further on its journey,
and this slide shows the final state.



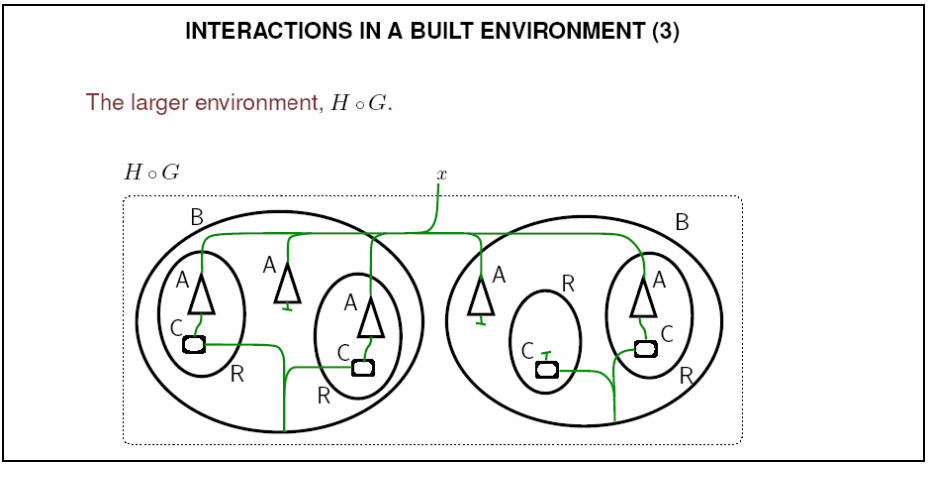

Can we use this very elementary kind of dynamics to explain other kinds of interactivity, possibly using different reaction rules? The idea is that we stick to the same kind of structure – bigraphs – but we use different reaction rules to describe what goes on in different kinds of system.



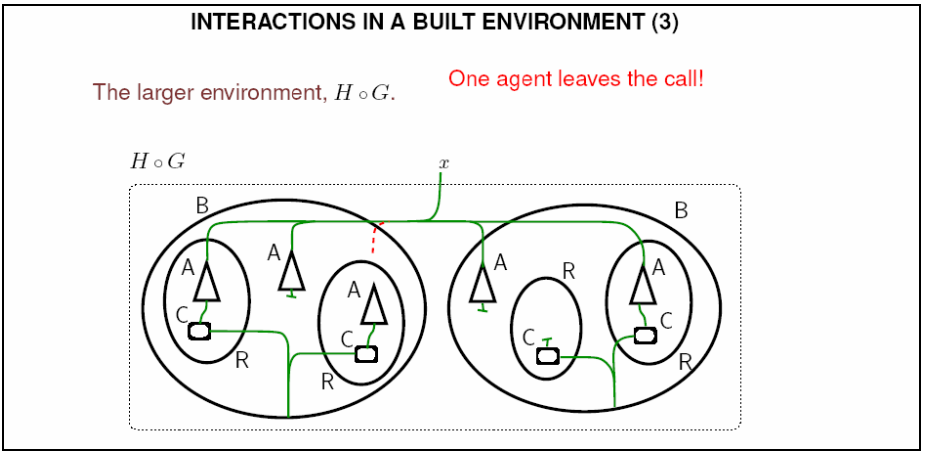

Look at an elementary, but nonetheless somewhat realistic, example of equipping a building with computers that are also sensors. I tried to model what happens when agents, who could be people carrying devices, move around the building. Here is a bigraph *G*, representing the system. **B** is a building, **A** is an agent, **C** a computer and **R** a room, so you have in the picture agents in rooms (and one not in a room but in the building, say in the corridor) connected to computers. The computers themselves are connected to the infrastructure of the building. This picture represents a subsystem of a larger system; it may be situated in a larger host system representing a larger piece of the world.



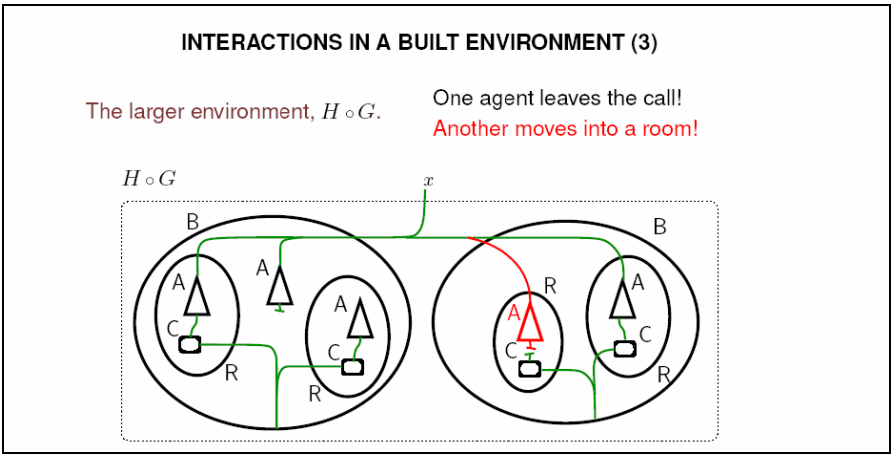

Here, at bottom left, is the system *G* again; you see that it has two parts (represented by the dotted squares) which may inhabit different parts of the larger host system *H*. The main picture shows the host system *H*, with two grey holes where the parts of *G* are to go. And when you put them in there, the wires of *G* must link up with those of *H* according to the letters *y* and *z*.




So here is the whole system, *H* with *G* inside it. This is a larger system, with another building involved. So there are actually two buildings side-by-side, with rooms in each, and agents and computers in each. What kinds of reactions can occur, representing activity in the system? First, think of the five agents **A**; most of them are in a room, but, whether they are in a room or not, they are having a conference call because that's what that green link represents. Various things can happen, and elementary reaction rules will represent tiny changes of state. For example, one agent may get bored and hang up on the conference call.

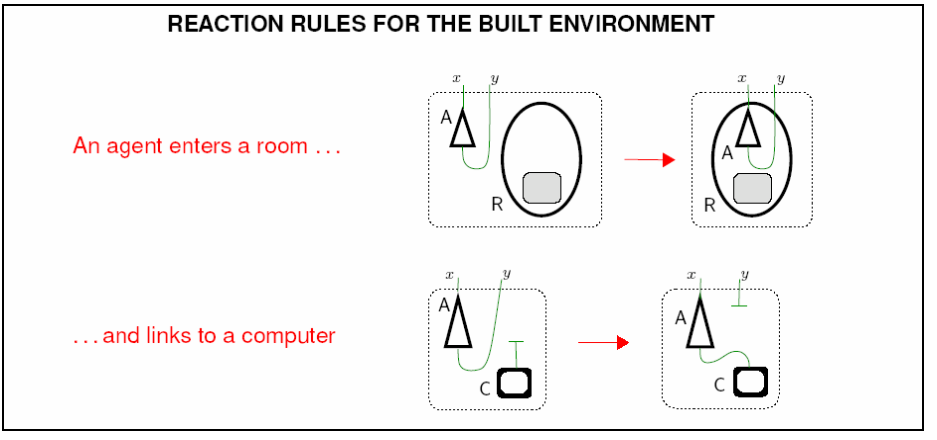


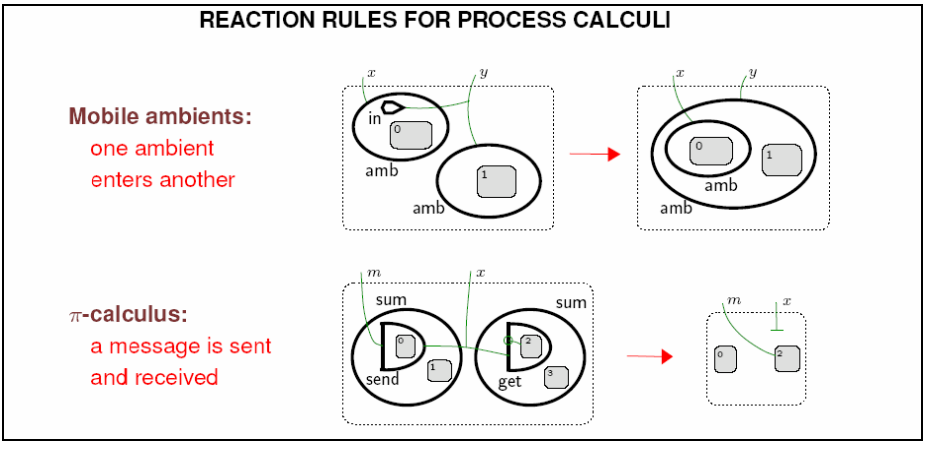

This slide shows the change; it's a change in connectivity. Next, one of the agents in a corridor may enter a room.




This slide shows the change, which is a change of locality. And once inside the room he may get linked to the computer, which has sensed his arrival.

To finish off this example, have a look at the reaction rules that make two of these changes happen. Take the top rule; the (red) arrow means you can change the shape on left of it to the shape on the right of it; so it represents an agent **A** entering a room **R** (but maintaining all her links unaltered). The second rule represents the computer **C** sensing an agent **A** and connecting with her. You can imagine any number of different rules, and you can wonder how many everyday events, including biological events (such as a cell ingesting a molecule) can be expressed by such rules. I shall not begin to answer that; but I would like to point out that the two calculi I mentioned, Pi calculus and Mobile Ambients, are fully expressed by this kind of rule.




In fact the top picture is a rule for Mobile Ambients; it represents one ambient entering another. And the second picture is a rule for the Pi calculus; it represents one agent passing a link called *m* to another. You can think of *m* as a link to a message. Some of the rules already used in biological modelling are similar. So there are chances that we are dealing with a universal model for reconfigurations of many kinds of discrete system.

Many questions can be asked immediately. What about hybrid discrete / continuous systems? Already there is evidence that these can inhabit the same kind of model, and therefore that differential equations can be part of the model. What about stochastic systems, where the reaction rules are equipped with likelihoods? These have already been incorporated with the Pi calculus in biological modelling, and indeed stochastic treatment is essential if we are to build models that actually perform, and allow us to predict real-life performance.

**CONCLUSION**

- The **challenge** is to devise computational *theories* for ubiquitous computing systems alongside the *engineering* of those systems...

**AND**

- ...the **sub-challenge** is to establish dialogue between the *theorists* and the *engineers*.

I have shown you a little of just one kind of theoretical model that can underpin the design of ubiquitous computing systems. The model focussed on mobility, but many other attributes need modelling, possibly in different ways. So, ambitiously, we hope for models that unite not only the computer engineering with computer science, but also computer science and biology.

That's the end of my talk.

**TWO LINKED GRAND CHALLENGES**

- **Science for the Global Ubiquitous Computer**
  moderated by Marta Kwiatkowska and Vladi Sassone.

- **Scalable Ubiquitous Computing Systems**
  moderated by Jon Crowcroft.

**Manifestos:** `www.nesc.ac.uk/esi/events/Grand_Challenges/`
**Discussion:** `wiki.science.luton.ac.uk/ubiqwiki/`

**UK UbiNet:** `www-dse.doc.ic.ac.uk/Projects/UbiNet/`
**EC FET initiative:** `www.cogs.susx.ac.uk/users/vs/gc2/gc2.pdf`

Here there are just a few bits of information about our Grand Challenge, which divides into two parts. I will leave them there during questions. Thank you very much.

# Interdisciplinary Innovation in International Initiatives

## Enabling Scientific Discovery via Disciplinary Interfaces: A Discussion Informed by the Trento Conference on Converging Science

John C. Wooley

Center for Research on Biosystems & Digitally enabled Genomic Medicine,
California Institute on Telecommunications and Information Technology,
9500 Gilman Drive, MC # 0043,
University of California, San Diego,
La Jolla CA 92093-0043

**Abstract.** Research in science and technology in the 21st century, emerging around the world in both academic and commercial settings, is characterized by the convergence of scientific disciplines and by the involvement of teams of investigators with diverse backgrounds. These developments reflect our increased knowledge about the complexity of nature, and correspondingly, the intellectual excitement that has grown at what were boundaries among disciplines. Over the past decade, deep, disciplinary expertise applied within multidisciplinary collaborations has been conducted on an ever larger scale, which has fostered many frontiers in interdisciplinary science. Scientific research at the interfaces - convergent science - presents significant challenges as well as opportunities: new organizational and technological approaches are necessary to enable the scale and scope of the human enterprise.

Examples of the research problems and the requisite intellectual skill sets and research environments have recently been articulated to summarize initial implementations of organizations exemplifying the opportunities of convergence and to provide a guide to action (Converging Sciences, 2005). We provide a further perspective on the context in which innovative efforts in converging sciences can successfully proceed; namely, the research will build upon a computing and information technology enabled research environment, that is, upon cyberinfrastructure. We explicate a specific early example of such a model, with novel organizational and technology features, that we hope can serve to inspire international efforts and enable comparable achievements.

## 1 Introduction: Leveraging on the Shoulders of Giants

All of the sciences have made remarkable progress over the past sixty years, and during this time, what was a small set of scientific disciplines has fractured into many domains or subdisciplines. The subdisciplines or fields have their own training programs, textbooks, journals, research goals, and so on, so each has acquired the properties of a discipline. The fracturing of the disciplines and subsequent growth of parallel professional infrastructure (for many fields) occurred in large part due to the many extensive advances in the methods and the technologies employed for experimental work, and the

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 14 – 37, 2006.

subsequent, rapid advances in our fundamental understanding of nature. As an example, the parallel scientific revolutions in the biological sciences and in computing and computer science and engineering are especially noteworthy. Both fields have essentially come of age and made a profound impact on the other sciences and on society over the past few decades; the significant, potential benefits from merging these two revolutions has been discussed elsewhere (for example, Wooley 1999) and is explored briefly below.

Much of this remarkable progress was driven by what is called a reductionist approach, in which investigators ask very well defined questions into specific attributes of natural phenomena. The questions are designed (in an attempt) to probe nature one step and only one step at a time, via a simple model system; that is, by its design, the research generally ensures the probe will produce an interpretable answer by using a narrowly focused approach that is of limited complexity and that provides for unambiguous experimental controls.

An important basis in establishing such an approach for scientific research in academic settings came in the USA from a study by Vannevar Bush (1945) that provided the basis for establishing the US National Science Foundation; namely, "Science: The Endless Frontier". Bush proposed a binary approach to research, in which industry only funds and conducts applied research, and academia conducts only basic research. To ensure progress and for science to serve society, government funding of basic science is clearly imperative under such conditions.

This model, over the subsequent years, has worked very well, although a gap between the funding of basic and of applied science came to exist; this gap is often termed "the valley of death" – the valley of death for many ideas or visions with practical significance but requiring more research for validation. In other words, little funding has been available from any source to test discoveries for their potential for practical application, other than seed moneys for proof of principle. Recent developments in university-industry partnerships may lead to solutions to this difficulty, as we will discuss below.

The diverging sciences, while so powerful in their approach to many questions, have now run into bottlenecks, particularly over the past decade, since nature does not follow explicit disciplines invented for human convenience and conducted through narrow or "stovepipe" organizations with equally narrow research objectives. The "energy" of science as a whole, of scientists engaged in getting answers to important questions, can not be contained by traditional policy expectations; that is, the ambition to solve these questions inevitably led to responses to solve the bottleneck. The responses are of three categories; namely, science policy shifts, changes in academic and industrial or commercial goals and the formation of new partnerships, and the appearance of novel interfaces among the many fields, known as interdisciplinary or converging science.

For an example of the first instance, science policy studies such as "The Pasteur Quadrant: Basic Science and Technology Innovation" (Stokes, 1997) have identified historical cases where the research boundaries established by and for the Endless Frontier did not make sense. (Science can have both basic and practical objectives, and even science done for practical objectives can have huge implications for basic understanding.) In the second category, investigators in academic and commercial sectors, during the past two decades, have come to share many research interests,

methods and objectives. In the third case, scientists have established partnerships of ever larger size, creating teams that had the necessary disciplinary expertise with an interest in a common research challenge requiring a converging sciences approach.

Well known examples for early steps toward convergence in the life sciences include the international Genome Project, the USA National Institutes of Health Protein Structure Initiative (or structural genomics), and the world-wide growth of bioinformatics and the hundreds of data repositories or resources, commonly called biological databases. Today, even greater convergence is happening with the rapid growth of computational biology and bioinformatics and the introduction of systems biology or an integrative, synthetic biological approach, to University research environments and educational programs.

The scientific revolutions of the 20th century, as described above in a particularly brief and quite simplistic form, served to position each scientific discipline upon an extraordinary vantage point; the conceptual vantage point itself can be considered a high ledge only reached via generations of discoveries. The ledge provides a clear view of the future paths for advancing fundamental knowledge and establishing profound technology applications. At the same time, to follow those paths, every branch of science now directly faces a remarkable challenge: addressing what has become known as complexity, the complexity of natural phenomena. The attempts to address complexity led to the bottlenecks that have made the introduction and implementation of converging sciences essential for future progress.

## 2 Converging Science: Science and Scientists Without Boundaries

Even as we enter the 21st century with enthusiasm and a revitalized sense of purpose for science and technology, the conduct of science itself, and not just the understanding that we have now achieved about nature, reflects complexity. Other scholarly endeavors can document the numerous root causes of the transformation within science. Instead, we describe, in this empirical contribution, an exemplar (within California, USA) for the emergence of translational science, with an infrastructure delivering convincing productivity (as viewed by citizenry, by those we need to underwrite the effort), and doing so throughout Science's Century even as we confront ever more difficult (complex) phenomena.

Both intellectual and economic pragmatism have accelerated the trend toward larger group efforts, or larger scale collaborations. The costs for research itself, such as contemporary advanced instrumentation, as well as the requirements for training a new generation, on one hand, and the nature – in simultaneous breadth and depth - of the expertise required to probe the challenges inherent in the complexity of nature, on the other hand, have led, in turn, to a distinctly different environment for the research community. Following the dictate that nothing succeeds like success, pragmatism has inspired partnerships (even among previous competitors) to emerge from the vigorous individualism of academic researchers and the robust rigor engendered through achieving commercial survival. Ever larger teams from diverse disciplines have arisen as senior individuals, previously distant – in expertise and often in geography - from each other, have been driven or inspired to collaborate.

The changing goals of science and technology reflect these twin pragmatisms and simultaneously, direct the creation of interdisciplinary collaborations. In particular, in

institutions situated around the world, bringing the products of basic as well as applied scientific endeavors to productive use has now become a pervasive commitment. Active researchers, science enablers or administrators, and the scientific and political leadership recognize the central importance of a reiterative or connected set of aspirations; namely, to advance scientific knowledge, facilitate science serving society, and ensure that society sustain science so that tomorrow's contributions can be even greater than those of today. While efforts by individual investigators within traditional disciplines will contribute directly and routinely, and will be required to achieve these goals, actual success will also require building teams of individuals with diverse expertise. Around the world, the ferment in science has already led to numerous informal or selectively funded partnerships, and to professional meetings aimed at stimulating interactions among disciplinary scientists and subsequent cross cutting collaborations.

As a step toward moving beyond the ideas of the Endless Frontier, the USA National Academy of Science, with inspiration from the leaders in biology and computing at the National Science Foundation, described over a decade ago how communication and computing technology could allow individuals from diverse backgrounds to work at distant institutions on common problems, or how distant laboratory efforts could be focused to address a shared problem through the use of computing technology (National Collaboratories: Applying Information Technology for Scientific Research [1993]).

An early instantiation of this idea - given in 1988 the new term collaboratories by Bill Wulf (to indicate both the power of interdisciplinary collaboration and the idea of joint laboratories collocated intellectually though physically distant) - began with telescience efforts – i.e., distant, sometimes very distant or remote, research partnerships and access to shared instruments facilitated by information rich communication –around the international sharing of remote instruments such as highly specialized, expensive and therefore rare, electron microscopes that provide novel modalities for experimental observation. Other implementations in medicine, chemistry and the environment soon followed.

The original telescience activity ( http://ncmir.ucsd.edu , http://ncmir.ucsd.edu; see also Tables 1 and 2), along with providing inspiration for applications in many areas, has itself continued to expand for well over a decade (Lee et al., 2003). Having reached maturity, these early efforts now form part of the "cyberinfrastructure" for converging science. The term "cyberinfrastructure" (CI) has been recently introduced by the USA National Science Foundation (NSF) to describe the integrated, ubiquitous, and increasingly pervasive application of high performance computing and advanced information technology (IT) approaches (Atkins, 2003). The novel features and the vast implications of CI for converging science, which similarly benefits and impacts society as well, are described below.

As governments and academic institutions move forward to obtain full advantage of the interactive, interdisciplinary, collective science research opportunities arising around the world, science opinion leaders need to remember and to emphasize that traditional disciplinary approaches remain as essential as ever. As a metaphor, this requirement is often described as maintaining the seed corn to grow next year's crop (another generation of scientists) and sustain scientific accomplishments. Therein lies

**Table 1**

| | | |
|---|---|---|
| National Biomedical Computing Resources | NBCR | http://www.nbcr.uscd.edu |
| National Center for Microscopy and Imaging Resources | NCMIR | http://ncmir.ucsd.edu<br>http://telescience.ucsd.edu |
| Biomedical Information Research Network | BIRN | http://www.nbirn.net |
| Joint Center for Structural Genomics | JCSG | http://www.jcsg.org |
| Protein DataBank | PDB | http://www.rscb.org |
| Wireless Internet Information System for Medical Response in Disasters | WIISARD | http://www.wiisard.org |
| Human Haplotype Project | HAP | http://www.calit2.net/compbio/hap |
| Lipid Metabolites and Pathway Strategy | L-MAPS | http://www.lipidmaps.org |
| Pacific Rim and Grid Middleware Assembly | PRAGMA | http://www.pragma-grid.net |
| Pacific RIM Experiences for Undergraduates | PRIME | http://www.prime.ucsd.edu |
| Optical Internet Protocol Computing | OptIPuter | http://www.optiputer.net |

the basis for advancing core knowledge, and the intellectual infrastructure for successful research training, upon which the interdisciplinary frontier science can build. In addition, while outside the scope of this article, for scientists to work without intellectual as well as geographic boundaries and to participate fully in international innovateions requires that the infrastructure to advance without boundaries as well; around the

**Table 2**

| | |
|---|---|
| NBCR | Advance biomedical research by developing tools to provide access to advanced cyberinfrastructure |
| NCMIR | Create the framework for end-to-end electron tomography |
| BIRN | Pioneer the use of grid infrastructure for medical research and patient care |
| JCSG | Automate each stage of structure determination and obtain structural coverage for all protein superfamilies |
| PDB | Provide the international community with persistent and searchable access to 3D macromolecular data |
| WIISARD | Bring cutting-edge wireless internet technologies from the hospital to the field treatment station |
| PRAGMA | Build sustained international collaborations and advance the use of grid technologies in applications |
| PRIME | Provide undergraduate students a research experience at a PRAGMA sites in Japan, Taiwan or Australia |
| HAP | Discover the genetic basis of human disease |
| LIPID MAPS | Develop interaction network of lipids and advance understanding of functional roles in cells |
| OptIPUTER | Implement powerful, distributed cyberinfrastructure to support data-intensive scientific research and collaboration |

entire world, the science opinion leaders and government science administrators will need to work together to ensure that such connections and level of transparency exist on an international scale.

The development of the internet and browser technology led to the most rapid impact ever to result from a new technology. Certainly, these developments have been a key aspect of the changed scientific infrastructure that enabled the rapid growth of distant partnerships among many individuals and accelerated overall scientific progress. The new infrastructure immediately changed one on one interaction between individual scientists in a very profound way, while also facilitating larger scale collaborations. The validation from the successes of many individual, small scale but novel, interactions is one aspect that has enabled the initial implementation of research teams within the context of converging science. As one result, telescience projects are extensively conducted and have even become routine.

The ease of access to the world's knowledge resources is another feature of the contemporary, changed environment for conducting scientific inquiry. Like society, science depends on an infrastructure. For research to proceed without intellectual boundaries and participate in international innovations requires that the pace of science be determined by creativity and not limited by infrastructure.

Underpinning the improvements in the scientific research environment has been the multiple decades of revolutionary changes in communication, computing and information technology. That the simultaneous advances on an exponential scale in computing, storage and communication - visible most notably in the Internet's penetration around the world - have led to an accelerated pace of change is widely recognized. (Obviously, information technology today drives all human endeavors, not just science.) A new term, now accepted into our vocabulary as cyberinfrastructure, was clearly needed to capture the novelty and extraordinary pace of impact on science and society arising from advances in the world of computing (as a whole). To an especially essential extent, the ability to engage in "Converging Science" and the potential for its impact requires and builds upon cyberinfrastructure. Before considering a specific instantiation of Converging Science, we need to consider the foundations for such novel, large scale efforts; namely, cyberinfrastructure.

## 3 Cyberinfrastructure – Multiplying Exponential Advances

The basis for the conduct of science and other activities within a society begins with an organization and structuring of resources such that individuals can create required products – that is, it depends on infrastructure. The extent of the utility of these numerous, organized and structures resources, along with their connections, provides the basis for progress, as well as the support for carrying out daily activities. The history of the transportation of physical goods, which first was done by water, then by rail, and then by major highways and the air, reflects the sequential establishment of new infrastructure that allowed major advances for transportation, just as the infrastructure provided by the printing press, and later the telegraph, the telephone, as well as radio and TV, extended the reach and subsequent utility of information and the empowerment of a larger fraction of the public. These classical achievements in infrastructure developed systematically but slowly.

The computer revolution, in contrast, has been driven by regular doubling of capacity of computing, storage and communication. These simultaneous advances on an exponential scale – more readily seen in the Internet's exception, unique rate of penetration around the world – have led to an easily recognized, accelerated pace of change. Along with improvements in information management, the infrastructure rooted in the Internet advanced by a factor of tens of millions in capacity in less than two decades, leading to profound changes in the information communication and access. The opportunities for science are correspondingly paralleled by those for our economy, culture, and so on.

Future advances in the Internet, or more generally, in cyberinfrastructure, will themselves be the result of Converging Sciences, rather than a single discipline. That

is, next generations in technology will themselves arise from continuing contributions from computing, physics, math and engineering, and will include integrating achievements in IT, nanotechnology, wired and wireless communication, micro engineered devices, photonics and numerous other technologies.

The central importance of cyberinfrastructure for Converging Sciences can be seen from the articulation of the NSF goals, by way of a Blue Ribbon Panel (Atkins, 2003), in this regard: (1) the provision of routine, remote access to instrumentation, computing cycles, data storage and other specialized resources; (2) the facilitation of novel collaborations that allow new, deep interdisciplinary and multidisciplinary research efforts among the most appropriate individuals at widely separated institutions to mature; (3) the powerful and ready access to major digital knowledge resources and other libraries of distilled information including the scientific literature; and (4) other features essential to contemporary research. Obviously, many of the historical roots of cyberinfrastructure began from the establishment of collaboratories, which first required and drove a centralized, intense application of cutting-edge scientific computing, advanced information technology, and tele-communication.

A meticulous consideration of the specific aspects of cyberinfrastructure in the life sciences is very relevant to this overview, given today's emphasis on using information technology to advance our understanding of biology, the details about a converging science model discussed below, and the similar exploratory efforts begun at Trento. However, each of these aspects has been presented elsewhere (see, for example, Arzberger et al., 2004b; Biomedical Computing, Visualization, Imaging and Informatics Resources, 2004), and thus, the specifics will not be discussed in depth in this article. An extensive discussion about the history of cyberinfrastructure (CI), and references to the early implementation of CI for the biological sciences, as well as a detailed analysis of the specific technology and science research issues and requirements can be found at http://research.calit2.net/cibio.

An excellent summary (CIBIO, 2004) of the specific goals for the life sciences describes "Building a Cyberinfrastructure for the Biological Sciences, which is available on the above web site as a PDF file, and as hard copy from the USA National Science Foundation. The considerations most relevant to this discussion are (1) the commitment to build a CI for BIO that meets specific needs in biological science research; (2) that the demands and potential are as significant for biology as for any other science; (3) that the interface between IT and biology is likely to lead to discoveries in areas that can not now even be imagined; and (4) that establishing the CI for the biological sciences will require a fully-international effort.

The stage for biological discovery is world wide, and the scientific advances in the biological sciences need to be shared world wide in order to accelerate fundamental discovery and applications for society. In turn, establishing major national and international partnerships may be challenging, but is an essential expectation on the world's science agencies. The first step involves a commitment to establishing and sustaining well organized, integrated, synthesized information or knowledge, and then to providing access to that knowledge to researchers everywhere.

## 4 Implementation of 21st Century Collaboratories

Four extraordinarily innovative, interdisciplinary institutes, each quite distinct in specific scientific goals but possessing a common vision for driving innovation and for delivering significant societal impacts, were established by the State of California in late 2000. These four institutes are collectively known as the California Institutes for Science and Innovation. Created with many similar organizational features, the institutes were founded by multiple academic and commercial partnerships, which extend in some cases even to national and international levels. In particular, the Institutes serve as test beds and models for the implementation of future, equally broad, collaborative research organizations, enjoying the digital, telescience world we have entered and addressing additional research objectives of comparable difficulty and scope (Inventing the Collaborative Research Environment for the Digital Future, 2005).

The convergence of sciences and the extent of divergent expertise within the Institutes for Science and Innovation are quite different from the very large scale collaborations (chiefly among physicists), frequently termed megascience, established at CERN in Europe or the Fermi or Jefferson Laboratories in the USA, and involving access to exceptional instrumentation and with more narrowly focused goals, albeit goals of ultrahigh challenge and importance for basic science. In contrast to the distributed research endeavors with many small scale projects, such as the Human Genome Project and the Protein Structure Initiative (Structural Genomics), the physical instantiation of each of the four Institutes will establish a close proximity of diverse researchers and will lead to synergy from the immediate access to resources and to the intellectual expertise of the whole Institute.

The California Institutes serve to interconnect investigators from multiple disciplines and from different institutions (campuses of the University of California or UC), into medium sized ("meso-scale") research groups aimed at the most highly challenging, contemporary problems. Each group of researchers includes a broad range of disciplinary expertise. The juxtaposition leads to the power to address far more comprehensive research problems than can be studied by conventional centers or through the efforts of many, traditional single investigator projects. The teams include not just academic participants (students, postdoctoral fellows, faculty, and academic professionals) but also experts from industry, government and the community. The origins, funding, nature of the interdisciplinary and disciplinary science contributions, and many other aspects of the four Institutes have much in common. To provide an explicit model for converging science considerations around the world, this paper will focus on one such institute, the California Institute for Telecommunication and Information Technology or Calit2 (although older references used the original acronym, Cal-(IT)^2).

## 5 Calit2: Converging Science to Extend the Reach of the Internet

A joint effort of the University of California, Irvine (UCI) and the University of California, San Diego (UCSD), Calit2 is formally an organized research unit (ORU), which in the University of California academic structure means, in essence, that it is an experiment in converging science that if successful will not only contribute

directly to the University's intellectual output, but will also inspire novel projects within more traditional departments and spawn additional, adventurous research thrusts. The governance and overarching vision of Calit2 resides at the Institute level, i.e., at the level of the partnership between the two UC campuses. At the same time, while the overall science programs are shared, progress on specific projects and effective impact is maximized by a campus-specific implementation of research, education, operations and industrial / commercial partnerships.

To provide the essential features succinctly and clearly, we discuss primarily the implementation at UCSD (Extending the Internet through the Physical World, 2001; Calit2@UCSD, 2005). The specific pre-existing research and education strengths, Centers and projects are complementary yet distinct (Inventing the Collaborative Research Environment for the Digital Future, 2005). However, the organizational and strategic implementations at UCI are similar in all essential aspects; indeed, most aspects are identical.

Calit2 represents is an institute of novel scale and objectives, driven by an exceptionally diverse and large team of investigators, pulled by ambitious, interconnected scientific grand challenges selected for potential impact, degree of difficulty, timeliness, and their ability to sustain synergistic contributions to the whole activity and to be enabled or pushed by common technology. As such, it provides an explicit, highly relevant model for the implementation of converging science in settings around the world.

The development of the Institute began with the convergence of more than 220 university faculty (and their postdoctoral fellows, graduate students and other investigators from diverse departments) and more than five dozen commercial partners. Following peer review of their proposal, the State of California provided funding ($100M USD total) for a high-tech building with the requisite instrumentation and research environment to be developed on each campus. The two state-of-the-art buildings occupy a total of 335,000 gross square feet and include numerous, highly specialized facilities, such as nanoscale fabrication labs, ultra-clean rooms for materials and communication research, and optical networking labs. The space is readily and highly configurable, to allow for innovative organizations that optimize interactions among people, their ideas and tools. A two fold matching amount was required (by the State) and obtained from contributions from private sources and industry, and a comparable amount is expected in the first five years from federal sources and Foundations. Among the investigators are computer scientists and engineers, mechanical engineers, mathematicians and statisticians, bioengineers, biological and biomedical scientists, chemists, materials scientists, physicists, environmental scientists, educators, and social scientists.

The unifying theme or essential core of the Institute involves advancing technologies transformed by the Internet; namely, linking advanced IT and communications to enable 21st century science. The Institute itself will be maintained as an experimental environment, in which the most exciting and appropriate projects that both drive and benefit from the broader vision will be housed for the time period necessary for their evaluation and implementation.

In conceiving Calit2, we took into account the exponential pace of advances in communication as well as computing, and the continuing fundamental transitions in

the Internet. First, scientific exploration and discovery, and many societal processes, will no longer be tethered to specific locations as digital wireless is applied and individuals will be contribute through mobile end points to the physical or fixed Internet. The end points will also include independent sensors and embedded processors, as well as the vehicles for human use such as computers and personal digital assistants. With science and society moving from conventional to broadband connectivity through fiber optics, and the actual optical technology advancing to multiple wavelength for encoding larger volumes of data, information will move ever faster among more participants; with the concomitant advances in optical technologies, the costs will decline. In sum, the goal of Calit2 is to provide test beds for specific applications serving to extend the Internet throughout the world. The test beds are conducted in integrated, flexible environments in order to respond to specific opportunities or difficulties; these are termed "living labs" (see below for a complete description).

## Organization of the California Institute of Telecommunication and Information Technology (Calit2)

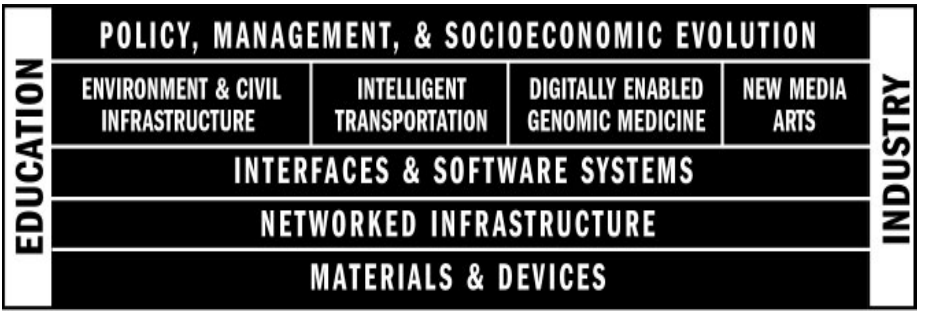



**Fig. 1.** A notional description of the interconnected, but distinct intellectual domains of the Institute, for which the applications layers interact and serve as research problem drivers for the technology layers, which provide the underpinning to enable the entire Institute; all of the layers interact in a bidirectional process with the education and industry pillars of the Institute. The Institute has an overriding sense of the important of policy, political, social and economic considerations for the state, and the need for quality management of the effort in toto. The considerations provide the top layer, for oversight, for finding opportunities and setting strategic directions for making a maximum impact, and for ascertaining how best to manage and sustain an effort on this unprecedented degree of size/scale, innovation, interdisciplinary scholarship, and local, regional and international impact.

Six application layers have been chosen for the test beds, under the criteria described above: environmental science, civil infrastructure, intelligent transportation, genomic medicine, new media arts, and educational practice. A core infrastructure established by three technology layers links the applications and provides the necessary foundation. These technology layers are materials and devices, networking, and interfaces and software systems. An overarching layer for policy, management and socioeconomic advances has been established to ensure the research is able to contribute directly to society, and two pillars, education and industry, interact with, contribute to, and benefit from all of the technology and application layers. The organization is shown in notional detail in Figure 1. [The figure is adopted from Extending the Internet through the Physical World, 2001. For more detailed descriptions of the vision, organization and implementation see also Calit2@UCSD, (2005); Collaborate, Innovate, Create (2005)).

Calit2 overall can thus be described as a set of research groups serving as interlocking layers. Beginning with the technology layers, the living laboratory for the Materials and Devices studies novel devices and materials that will enable tomorrow's Internet infrastructure. (The research includes such aspects as molecular materials, optical, wireless, and storage components; and micro-electro-mechanical devices (MEMS). Networked Infrastructure explores the path to advanced communication. (The research includes digital wireless, broadband communication, architecture, protocols and management of networks, photonics, sensors and storage.) The Interfaces and Software Systems layer studies the requisite informatics and computing aspects for extending the internet. (The research includes software for secure, scalable distributed systems, mobile communication and computing agents, and information technology of knowledge management and data mining, and also human-computer interfaces.)

Next, moving to the applications layers, each of these was chosen both for their dynamic potential for impacting and being impacted by the overall Calit2 infrastructure and the technology layers (as described above), and for their potential for improving the quality of life, with immediate application for California and the large market segments of the State's economy that will be transformed by the future Internet.

Education and industry represent teams whose impact and contributions are pervasive among the societal, scientific and technology objectives of Calit2, and thus they can be envisioned as vertical layers adjacent to all of the other layers. Each application and technology layer will contribute to new educational processes, and each will be influenced by new generations of students and their experiences, suggestions, and accomplishments. Similarly, industry has numerous ways to contribute to the academic efforts within each layer, including education. In turn, each layer will contribute to the objectives of our industrial partners, who will help train some of our young scientists, and some of these scientists will ultimately join our industrial partners.

Another goal of the Institute is to create a dynamic collection of the living laboratories of the future, or simply, living labs, which are dynamic entities created at the right time to include the right individuals with the right expertise and tools (and interpersonal skills). The strength of the interactions with the regional community and with industry allows Calit2 to establish the labs not just within the physical space of the buildings but across both campuses and across all of southern California. Research

prototypes from academia and early implementations of products from industry are brought into the Calit2 environment, creating the living laboratories, which will be sustained as long as a strong synergistic value exists. The living labs first serve to build and test integrated systems under real world conditions, and consequently, to deliver exceptional insights. For academics, the labs provide insight into the most exciting domains for roughly the next few years, whereas for industrial partners, the living labs provide insight into future mass markets up to roughly five years before the markets themselves exist. Academic researchers evaluate software and experimental advances, while industrial colleagues can evaluate the potential of prototypes and see what the technology might enable for in the way of future products. At the same time, the science and technology activities are conducted in a cultural context, and a range of scholars, including those from business management, cognitive science, policy, and education, use the living labs to explore another dimension; namely, topics such as the human potentials for creativity, training and learning, and productivity.

The Institute is also distinct from those large scale research and/or service organizations that are commonly called centers. In particular, the Institute is characterized by prototyping research environments, advancing basic science in the relevant domains, testing future technology implementations, and providing a path for future instruments and methods. These activities are conducted in concert with industrial partners, who invest thereby in their own future. Besides the differences in scale of effort and the presence of the entire range from basic discovery to application test beds within the Institute, university centers in general are more narrowly focused, and less involved in advancing research infrastructure or exploring new research environments.

Discussing the nature of Calit2 requires asking how such an Institute compares with extant large scale research organizations within Universities, which, like Calit2, in the University of California system are given special status as organized research units (ORUs), reflecting novelty, fusion of ideas, and/or the breadth of faculty involvement. Many traditional university centers focus on a common intellectual theme (for example, at UCSD, these include Molecular Genetics, Cancer Research, Molecular Medicine, Immunology, and Biological Structure). Others, such as the High Performance Computing Centers (under various names over the past twenty years) established by NSF and sustained as resources for the Nation since 1985, have the capacity to deliver services required by the community and have a goal of doing so in an outstanding way into the indefinite future (if given the capability through organizational excellence and adequate funding). Such is the relationship between Calit2 and the San Diego Supercomputer Center (SDSC), which delivers scientific compute cycles to the Nation in a rich-support-infrastructure, while in parallel conducting selected, specialized scientific and technological research within the domains of scientific computing. Calit2 does not commercialize products or deliver such services, but rather anticipates the future and enables continued excellence by organizations like the SDSC.

This type of science faces another hurdle not so common to conventional academic departments; namely, the need for recruit, reward and retain, academic professionals, who are technical experts more experienced than academic apprentices – students and postdoctoral fellows – but who are generally not engaged in the teaching and

administrative business of the university. Rather, they are totally focused on the science at hand, as would have been possible otherwise only in industry. These academic professionals -among other contributions - serve to connect academic and industrial research efforts within the living labs. The value these individuals find in the Calit2 adventure and their corresponding tenure and degree of engagement will certainly play an important role in the ultimate level of success and impact of the Institute for California, the Nation and the world, and its impact on the ways in which science is conducted.

## 6 Innovating a Biomedical Layer Bridging Engineering and Biology

To appreciate why a biological and biomedical layer became an obvious and necessary layer with Calit2 - in terms of its ability to benefit from IT and the advances in the Internet and in telecommunications, and its ability to drive the technologies as well - consider the overall environment for 21st century biology, which can be seen metaphorically as the setting for a grand opera performance. The early, extraordinary and even unanticipated progress and subsequent success of the Human Genome Project opened a vast panoply of difficult science research problems and important societal objectives. To move further requires a considerable amount of highly difficult experimental biology research, including on the specific information content within the genome, and the genome's read out in RNA, and in turn, from the messenger RNA species, its ultimate product, the proteins. The world of proteins, loosely speaking, can be considered as a second phase for genome scale science, a science now termed proteomics, which will be characterized by even larger data sets than those from genomics. Besides working out a parts list for living systems and establishing a deeper understand of fundamental protein chemistry, the challenge for some time to come is to probe, characterize and understand a set of interconnected webs of biology that collectively enable, sustain and direct the processes of life; namely, the metabolic networks, genetic circuits, and signaling cascades.

While none of these research problems are easy or will be solved in the near term, the challenges have accelerated the implementation of novel organizational and intellectual approaches for biological science research, as well as the rapid entry of computing technology into biology, in order to store, manage, query and mine the highly complex, hierarchical, multiscale nature of biological science experimental data and to begin to model how biological systems work. The twin revolutions in computer science and technology and in molecular biology have now become ideal for each other. Indeed, the exponential growth of computing technologies and of biological data has driven the establishment of numerous frontiers at the interface of biology and computing (ref NRC studies, 2005).

A particularly exciting and powerful application domain within 21st Century Biology is translational medicine, the use of the knowledge from the genome project (and its extensions into proteomics and systems biology) and of medical bio-informatics to advance the application of molecular approaches for understanding human biology and delivering health care. Translational medicine, or bringing science from the bench to the bedside, will depend critically on organizations like

calit2 that provide the infrastructure for research at a frontier between computing and biology. The components of such an approach are commonly called personalized medicine, predictive medicine and preventive medicine, leading to one of the metaphorical stretch goals of Calit2 to deliver the products of the genome for improved health care by 3PM! This is genomic medicine.

Similarly, wireless devices for physical fitness training and physiological monitoring were early implementation of digital monitoring and communication technology (for health or wellness care through healthy citizenry). Digital biocompatible sensors will go beyond today's monitors of gross physiological parameters to measure human body chemistry, traditional chemical and metabolic panels for patient monitoring and care, and will do so 24 by 7, as our bodies will truly go on line. An early implementation of this is the video pill, produced and used in Israel and Japan, which allows non-invasive imaging of the GI system. Imagine following glucose or insulin levels, or white cell count, remotely, for outpatients at risk. Microarray analysis of human cell properties in health and disease is another example of high throughput large and complex data sets that will be increasingly available and applied to health care; so are multiscale, multimodality analysis of human health and disease via MRI, PET, CT and other technologies. These all represent digitally enabled applications, implemented through intense computational technology.

Out of these exceptionally converging sciences, Digitally enabled Genomic Medicine or DeGeM was born, that is, the neologism DeGeM was introduced as a key layer with calit2. Its implementation bridges people, tools and their ideas arising from the core disciplines. Specifically, DeGeM is a rich intellectual fermentation engaging efforts from biological sciences (including chemistry, biochemistry, biophysics, bioengineering, bioinformatics, computational biology, neuroscience, cell biology, molecular biology, evolutionary biology), medicine (including biomedicine, molecular medicine, translational medicine, radiology, health physics, clinical trials and practice) and engineering (computer science and engineering, mathematics and statistics, high performance, grid and cluster computing, internet and telecommunication technology, database and data federation and other aspects of information technology, sensors and sensor nets), with assistance from other layers notably including materials science but with some overlaps with environmental science and even intelligent transportation with smart vehicles that not only monitor traffic but also communicate health status parameters bidirectionally).

## 7 Digitally Enabled Genomic Medicine (DeGeM of Calit2)

As a first step, DeGem was built around a set of shared principles about the path for the biosciences in the future and about the requisite informatics and computing, and the use of multiple modalities while working on multiscale information about biology (DeGeM, 2004). These principles are (1) Research in 21st Century Biosciences requires a comprehensive Cyberinfrastructure; (2) the technologies and the level of biological understanding are advanced enough today that it is now actually possible and also absolutely necessary to facilitate the convergence of disciplines and build frontier research at the interface between computing and biology; (3) the key biological projects

within DeGeM should focus on the top requirements for advancing biomedicine and biology; namely, multi-scale experimentation, informatics and modeling; multimodal, computational and data mining efforts to understand the vast scales of time, space and organization of biological systems; research on dynamic form and function of biological macromolecules; (4) the maturity of experimental approaches and of scientific computing for the biological sciences and biomedicine provides the basis for robust generation of theories helping to explicate a synthetic, integrated view of biology.

The fourth point requires more explanation. Theory, let alone modeling, simulation and computational analyses in general, has been actively neglected within the research strategies of biological sciences, which may well have been a necessity during the reductionist era of rapid discoveries in the biological sciences. The era was characterized by the rapidity with which exceptional insights followed every pronouncement that biology was now well understood, and no new challenges remained. Today, however, given the parallel revolutions in biology and computing, for the first time, it has become possible to build a theoretical perspective that integrates the various approaches (notably, observational and interventional experiments, modeling and simulation, pattern recognition and data mining, discovery or exploratory science and hypothesis driven, reductionist studies), and link experimental observations by different modalities on multiple scales, and in doing so (in bringing these still separate threads of research together) leads to a synthetic, integrative biology.

The computing and IT applications within DeGeM will be selected and implemented when need and value has been demonstrated in the context of cyberinfrastructure. These layer-specific applications are Seamless Capability and Capacity Computing; Innovation, Development and Implementation for Instruments & Experimentation; Data, Standards and Database Management, Analysis, Integration; and Computational Modeling & Simulation.

A collection of previously extant UCSD federally funded research centers and resources, which have interacted and collaborated routinely and extensively, originally provided the foundation and drove the trajectory for developing DeGeM as the biomedical layer within Calit2. These centers (see Tables 1 and 2 for more details, including the overarching vision and goals, and the relevant Universal Resource Locator, URL, to the web site) include the National Biomedical Center Resource (NBCR), the National Center for Microscopy and Imaging Research (NCMIR), the Joint Center for Structural Genomics (JCSG), the Protein Databank (PDB, or Research Collaboratory for Structural Biology, RSCB), and the Wireless Internet Information System for Medical Response in Disasters (WIISARD). NBCR, NCMIR, JCSG were established and integrated in the context of calit2 through the sponsorship of another organized research unit of UCSD; namely, the Center for Research on Biological Structure, now termed the Center for Research on BioSystems (CRBS); CRBS served as a model in establishing the organization and many aspects of the scientific focus of DeGeM, and CRBS remains a central hub in the development and deployment of DeGeM.

Like the larger Institute (Calit2), both individual investigator projects and larger scale group projects, and even an organized research unit interconnecting some of the existing centers, contributed to DeGeM's foundation and growth through establishing

a novel vision capturing the excitement in modern biology, initiating full scale research projects, obtaining funding, and so on. Thus, there is a nested, interlinked set of larger scale, center-type entities that provided the foundation and now will interact within DeGeM with the entire calit2. All of the layers of calit2 have some of these interlinked and nested attributes, which allowed the Institute to begin to function immediately upon creation and work productively in existing space even before the new buildings were completed. At the same time, new projects of the scale of Centers began within the intellectual framework of DeGeM and continue to advance rapidly (DeGeM, 2004). These include the Human Haplotype / Human Genetic Disease Project (or the HAP Project, see Table 1 and http://www.calit2.net/compbio/hap), the Lipid Metabolites and Pathway Strategy (Lipid MAPS; http://www.lipidmaps.org), the Pacific Rim and Grid Middleware Assembly (PRAGMA, 2004) and its undergraduate research component, the Pacific RIM Experiences for undergraduates (PRIME).

The OptIPUTER project, a component of three different layers within Calit2, is closely interconnected with other on going research projects within DeGeM. The OptIPUTER has tested and is now building a powerful, distributed cyber-infrastructure to support data-intensive scientific research and collaboration (Calit2@UCSD, 2005). Another key aspect of the diverse activities in which DeGeM is engaged is a new, campus-wide, interdisciplinary Bioinformatics Graduate Education Program, which brings together investigators from physics, chemistry-biochemistry, mathematics, various biology specialties, pharmacology and other biomedical domains, bioengineering and computer science to create and sustain the program, and whose goal is to prepare the next generation of biomedical researchers through rigorous exposure to both experimental and bioinformatics research (http://bioinformatics.ucsd.edu). Other previously funded activities beginning to be interconnected and expand within DeGeM include the large scale modeling of biological signaling and regulatory pathways (Cytoscape: http://www.cytoscape.org/), multiscale (molecular to organ) mechanics and dynamics of the normal and diseased heart (within UCSD's Cardiome lab), via the biomechanics of cardiac myocyte and matrix remodeling and cardiac electromechanical interactions (http://cardiome.ucsd.edu and Continuity: http://www.continuity.ucsd.edu/), and other projects in bioinformatics, computational biology and computational biomedicine (see DeGeM, 2004).

Each ongoing group project within DeGeM has many common elements, goals and methods with the others, and they intersect, interact and collaborate in a wide variety of ways. Most notably, the universal themes (those present in every instance) among all of the projects are building and advancing the cyberinfrastructure necessary for the biological sciences in order to enable novel directions and the explicit application of scientific computing and information technology to accelerate progress in addressing major biological science questions. Along with information technology, all of these projects bring together some aspects of the physical sciences (physics, biophysics, chemistry, and biochemistry), quantitative sciences (mathematics, statistics and computer science), engineering (engineering technologies and bioengineering). Each project also contributes directly to advancing biomedical science.

Table 2 provides a brief overview of the goals and technologies of the major, original interdisciplinary Centers and other large scale, multi-investigator projects, equally interdisciplinary, from which DeGeM was implemented. The breadth of the

goals - as given briefly for each activity in this Table - shows the extent of converging science that was already an inherent component of each of these initial, pre-calit2 efforts; as a consequence, each of these activities will serve to benefit even more from the impact of converging science in being brought together in one location, and encouraged to engage in active discussions, coordination and expanded collaborations through a singular, ambitious vision. That is, the nature of the DeGeM component activities means that their presence within Calit2 will provide many advantages for the Institute and the activities. Having a coherent physical location and a formal intellectual intent to build ever stronger and wider bridges among the activities (to facilitate communication, cooperation and collaboration) will lead these projects into new directions and accelerate their progress.

## 8 Early Implementations of Living Labs for DeGeM

Four components from the centers and multi-investigator projects that provided the foundations for DeGeM have already become Living Labs and are being extended in major ways (Calit2@UCSD, 2005); others will follow. Having proceeded quickly and been most fully implemented, two application-driven living laboratories are flagship projects for Calit2 and particularly representative of the Calit2 expectations and goals for Living Labs. These are the Biomedical Informatics Research Network (BIRN) and the Wireless Internet Information System for Medical Response to Disasters (WIISARD). Two other living laboratories are newly developed, rapidly expanding, and already well interconnected to the other components of the Institute; namely, the HAP Webserver and the Visible Cell Project.

BIRN was established (in 2001) to support both basic and applied biomedical research (i.e., to advance biomedicine from studies on fundamental knowledge of biosystems to studies on translational and clinical research), and serve as a model for the introduction of cyberinfrastructure into clinical practice as well as into biomedical research (Lee et al. 2003; Peltier et al., 2003; DeGeM, 2004). Using diverse, multiscale neuroscience observations for this initial test bed, the objective of BIRN is to build and extend the software and protocols necessary to share data (and its interpreted implementation, information and knowledge), to analyze and mine data, and to establish the hardware most appropriate for the BIRN teams around the entire National Collaboratory.

The two major pillars on which BIRN was erected, that is, NCMIR Telescience and NBCR, which are within CRBS at UCSD, have contributed extensively to the rapid advancement and initial deployment of a scalable cyberinfrastructure for the biological sciences (CIBIO, 2005). As every aspect of computing has advanced and become more powerful for science, their impact became international in scale. These efforts consider the range of details needed for effective and productive implementation of cyberinfrastructure, which in toto includes high bandwidth, advanced communication networks, federated, distributed information repositories, scientific computing facilities, and software engineering. They also involve applications that are being brought together into a common technology to address the requirements and growing potential of collaborative test beds and their larger scale implementation as full

collaboratories for multiscale investigations in biology (Biomedical Computing, Visualization, Imaging and Informatics Resources, 2004).

The investigators around the Nation who are brought together through BIRN are able to address research into diseases of higher complexity than individual investigators could study. The underlying basis for the success of a test bed or ultimately, of collaboratories, depends on bringing together a wide range of the relevant domain expertise, providing access to state of the art instrumentation of the appropriate range of modalities, unique support facilities and resources, and for providing overall a platform that allows a comprehensive focus on the relevant application.

The cyberinfrastructure for BIRN was established in an unusual yet highly productive way. From the very beginning of the project, a coordinating center (at UCSD) was established to ensure effective, rapid management decisions of the highest technical standards made consistent with overall objectives. The coordinating center set the hardware, software and communication standards, ensured agreement on common scientific and technological goals, and linked the extant data of all of the individual regional centers together. Far more frequently, in large scale progress, distributed authority has interfered with effective science; the USA National Institutes of Health National Center for Research Resources deserves considerable credit for recognizing the importance of effective coordination for this project in order to ensure rapid progress and the establishment of an effective model for other efforts in converging science and for its applications to clinical medicine.

Along with the coordinating center, BIRN began with from three initial research test beds; namely, the Function BIRN, the Morphometry BIRN, and the Mouse BIRN. The Function BIRN is composed of 11 Universities studying human brain dysfunctions related to schizophrenia; Morphometry BIRN, which is composed of nine research institutions investigating structural differences in the human brain to see if these can serve to categorize for diagnostic purposes a set of human brain diseases; namely, unipolar depression, mild Alzheimer's disease, and mild cognitive impairment; and Mouse BIRN, which is composed of four institutions using mouse models of human disease at a range of anatomical levels or scales in order to test contemporary hypotheses concerning a set of human neurological disorders; namely, schizophrenia, attention-deficiency hyperactivity disorder, multiple sclerosis and Parkinson's disease.

BIRN has also provided such important advances as normalizing the images obtained from different modalities, providing a common set of platforms for data sharing, transforming collaborations among the participating institutions, enriching grid technologies by proving a new set of domains, and more generally, provided models for the establishment of similar collaboratories for environmental health, for response to natural disasters and disease outbreaks, and for other organs and diseases. In Calit2, the SDSC Knowledge and Data Engineering Lab, the Lambda Grid Living Lab, the Interfaces and Software Systems Layer, and the OptIPUTER Project will all serve to leverage an increased productivity by BIRN.

The HAP Webserver was established to facilitate the analysis of human conditions and diseases that are likely to result from the interplay of multiple genetic and environmental factors. This server established a means for the readily achieved computation to predict genetic variation on each human chromosome. This differs from the Human Genome Project and other high throughput sequencing efforts, which

obtain only the joint information, that is, from both chromosomes. Genotype data is directly input from anywhere in the world into the Hap Webserver, which computes phasing information on haplotype and establishes the blocks of limited genetic diversity.

The other newly established DeGeM Living Lab, the Visible Cell Project, aims at advancing our knowledge of very large cells, such as the "extreme cells" of the nervous and muscular systems, by bringing together multiscale data collected through diverse modalities (calit2@UCSD, 2005). Structural details are being collected at all levels, from micro or atomic, to macro or anatomical levels, and these architectural features are connected to dynamic or functional information. A unique multiscale, federated database has been built and is being extended to capture, integrate, and store the information and provide a platform for sophisticated data mining and multiscale modeling.

WIISARD (initiated in 2003) is the other early, flagship Living Lab for DeGeM and Calit2. Established after analyzing the risks and consequences from events like the 9/11 terrorist attack and the related need to have more effective responses to natural disasters, WIISARD has initially focused on how best to enable and protect the efforts of first responders (Calit2@UCSD, 2005). This includes an urgent need to establish a wide, diverse range of two-way information communication with those responders. The information could concern either natural or human created accidents or major events, including terrorist actions (which might involve chemical and/or biological agents, and potentially, small nuclear devices). Nothing short of extreme measures are certainly necessary to deal effectively with the level of urgent, immediate, but careful or intelligent, response required, since at that level of impact on civil infrastructure, the individual health care facilities will be unable to obtain information readily, and therefore, unable to respond adequately. In worst case scenarios, there will be significantly increased risk to human lives.

WIISARD builds upon the extant breadth of clinical and other health care knowledge and uses the Calit2 expertise with wireless Internet IT. Consequently, WIISARD has established the first method for actual deployment at multiple terror or natural disaster sites. Subsequently, WIISARD has tested and implemented the technology to provide life support for a large numbers of victims on a time scale of hours to even days. In addition, the wireless Internet IT technology implementation being established provides the means to communicate an exceptionally detailed assessment of the conditions and situation as an assessment so exceptionally detailed as to allow a National level response.

WIISARD is a local implementation involving first responders from the County of San Diego Metropolitan Medical Strike Team (MMST), which is a civil interdisciplinary activity, including public health, paramedic and fire officials, law enforcement personnel, and experts in the management of hazardous materials), working with Calit2 computer scientists, information scientists and a variety of engineers and with UCSD's School of Medicine physicians. The activity began with pilot efforts and is building toward ever larger simulations to test current capacity and evaluate options necessary to meet increasing levels of threat. For example, a simulation of "dirty bomb" explosions in commercial establishment, the largest civilian emergency response drill ever in the County, was conducted (in 2004) and others are planned.

Overall, WIISARD is a Calit2 integrated application bringing the power of state of the art wireless technology to facilitate clinical settings from hospitals to ambulances or even field treatments. WIISARD provides the ability to increase the situational awareness for first responders, to record appropriate medical data, to monitor patients at high risk and to do so under difficult conditions, and to enable two directional communications with local hospitals to obtain and update relevant medical data.

WIISARD has been established for the entire southern California region by way of the partnership of the UC Irvine and UCSD divisions of Calit2. For the broader, long term goals of DeGeM, WIISARD will provide insight into how to extend health care in clinical settings and serve to drive advances in the security and robustness of wireless communication. WIISARD complements the kinds of contemporary activities delivered by the fitness industry and the efforts to build a stronger IT support system for health care in our Nation's hospitals. It also provides an approach to the novel use of advanced, scalable cyberinfrastructure, developed through BIRN and other DeGeM activities, to improve health care and contain costs.

More generally, the research by BIRN, WIISARD, the HAP Webserver, the Visible Cell Project, and other DeGeM activities, will be followed and when possible, extended to health and wellness care in clinical and home settings. In clinical settings, the studies will be translated to facilitate the search for the epidemiology of infectious disease, improved applications for earlier diagnosis, status monitoring, and the evaluation of prognosis. In home settings, the studies will be incorporated along with sensor net and communication research to provide for safer home care (in an era requiring shorter hospitalization and minimized bed count), in part by extensions of the popular methods for tracking physiological status and fitness parameters, which were the first wireless wellness monitors.

## 9 Implications and Conclusions

The Calit2 buildings are done; they are populated at UCI and are now being populated at UCSD. The Living Labs, the Layers, and the overall activities of Calit2 have been proceeding from the origins of Calit2, without waiting for the acceleration from a common physical instantiation. Nonetheless, the work of the Institute, like other converging science efforts at any scale, has just begun. On a regional level, that of the State of California, the four Science and Innovations Institutes will complement and might interconnect or collaborate from time to time with the UC Discovery Industry-University Partnerships, which are more focused and smaller in scale individually, but collectively present an intriguing interface.

In observing and evaluating the implementation of converging science, more generally, it will be exciting not only to complete the implementation at Calit2 and be engaged in its unique contributions and mission, but also to follow, interact with, benefit, and influence parallel scientific and innovation activities at various levels of effort around the world. For example, a parallel, early phase implementation (along the frontiers at the interface of computing and biology) has begun with the interlocking partnerships of the local and Italian federal government agencies, academia (University of Trento) and industry (Microsoft Research) in Italy. Numerous other examples of converging science in European settings are described in the "Converging Sciences."

In Asia, highly interdisciplinary, large scale collaborations in Singapore have been assembled with close proximity at two clusters of buildings, the Biopolis and the Science Park, sometimes called the Techopolis (for fundamental biology and for commercial groups, respectively). The opportunity was created by their government's commitment – specifically, by the Agency for Science and Technology Advanced Research, A*STAR - to the value of innovation and discovery catalyzed by converging science (Cyranoski, 2005). These clusters bring into close proximity a productive set of extant Centers and other large scale research and development efforts, which are more narrowly focused.

The creation of these coherent adventures in convergent science, in the USA, Europe and Asia, reflect similar motivations in the inclusion of both academic and industrial partners and both basic and applied research (or actually, the inclusion of the continuum from basic to applied research). These efforts around the world are of comparable scale, and all of us should watch carefully watch and absorb the lessons from the sibling Institutes (whatever the actual name of the effort).

Disciplines, specifically in the form of university departments, will continue and are needed in the era of converging science; no one should see converging science as being in conflict or even contrast to disciplinary science. An appreciation of the complementarity rather than conflict is imperative. The disciplines themselves will flourish, that is, they will continue to grow, provide important educational and science roles, transform or morph, and address still to be established objectives while feeding into and upon the convergent science practiced in unique infrastructure environments. Not only will many frontiers at the interface of disciplines be explored and exploited through scientific convergence, the convergence of aspirations and practices for industry and academia will accelerate.

With regard to the future of converging sciences and their ultimate impact on research and on society, two final considerations seem most relevant. First, as Wally Gilbert (1991) presciently pointed out in his article (on a specific instantiation of converging science: the introduction of bioinformatics as a routine research tool and equally routine access to web based information resources for biology), entitled "Toward a New Paradigm for Molecular Biology," scientists need not fear becoming obsolete due to the introduction of new technologies and approaches to conducting science, but rather the establishment of new technologies and the contributions from scientific convergence will accelerate progress on fundamental disciplinary questions, enrich educational opportunities, maintain the health of the disciplines themselves, and establish a true translational science serving society. Second, our institutions, government funding agencies, and individual scientists should sustain the importance of reinvention and subsequent invigoration, as described in poetic form by Bob Dylan (1965): "He not busy being born is being dying."

## Acknowledgements

There might never have been any California Institutes of Science and Innovation, without the prescient efforts by Erich Bloch, Alan Leshner, David Kingsbury, Rich Nicolson, and other executives at the National Science Foundation in the 1980s, who build upon other early models, such as Engineering Centers, to establish the NSF's

Centers in Science and Technology, which were radically interdisciplinary in their day, and they proved to be incredibly innovative and as such, become the model for the State of California in envisioning and even reviewing its Institutes of Science and Innovation. The conceptual foundation for the California Institute of Telecommunication and Information Technology arise through the remarkable vision of numerous individuals; in particular, Dick Atkinson, Bob Dynes, Marsha Chandler, Ralph Cicerone and Marye Anne Fox played important roles in establishing, developing, implementing and extending the highest level of vision for Calit2, while the specific attributes and components were innovated by Larry Smarr, who also, in conjunction with Bob Conn, took a lead role in ensuring that enough funding was available to start the engine, and who has since worked with the 200 plus faculty, including our many dedicated and inspired colleagues at UC Irvine, to enable the continuing adventure on the Torrey Pines Mesa. Ramesh Rao and numerous layer leaders and other faculty participants got the right operations management in place and established the specific details for the journey of innovation.

Influenced by inspirational discussions with Larry Smarr and Mark Ellisman around the vision for Calit2 and their goal to include an innovative biomedical and preclinical component, as well as by my preliminary plans for introducing a translational and genomic medicine course at the UCSD School of Medicine, the author chose the name DeGeM for the reasons described in the text above, although the reader will also recognize the sense of The Gem for the Institute in delivering on the promise to the people of California, the Nation and the World. The projects and the many other individuals involved in its implementation can be found in the DeGeM Brochure given in the references. DeGeM also owes much of its vitality and structure to an extant organized research unit of UCSD, the Center for Research on Biological Structure, now termed the Center for Research on Biosystems (CRBS), led by Mark Ellisman. In the early stages of applying IT in radical new ways, the thinking of Bill Wulf and David Kingsbury both provided a basis for subsequent cyberinfrastructure developments and extensively influenced my thinking and my choice to become involved ever more deeply engaged in the interface between computing and biology, and to begin to consider the challenges and opportunities provided by converging sciences.

## Bibliography

Arzberger, P., Papadopoulos, P. (2004a), "PRAGMA: A Community-Based Approach to Using the Grid." Access Grid Focus 2: June 2004.

Arzberger, P.W., Farazdel, A., Konagaya, A., Ang, L., Shimojo, S., and Stevens, R.L. (2004b), "Life Sciences and Cyberinfrastructure: Dual and Interacting Revolutions that will drive Future Science." New Generation Computing 22: February 2004.

Atkins, D. (2003), Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Computer and Information Science and Engineering Directorate, National Science Foundation, Arlington, Virginia.

Biomedical Computing, Visualization, Imaging and Informatics Resources (2004), National Center for Research Resources, National Institutes of Health, Bethesda, Maryland.

Bush, V. (1945), Science: the Endless Frontier. National Science Foundation, Arlington, Virginia.

Calit2@UCSD (2005), University of California, San Diego Division of the California Institute for Telecommunications and Information Technology. La Jolla, California.

CIBIO (2004), A BIO Advisory Committee Workshop, July 14-15, 2003: Building a Cyberinfrastructure for the Biological Sciences." National Science Foundation, Arlington, Virginia.

Collaborate, Innovate, Create (2005), Calit2@UCSD Building Dedication, Calit2, UC San Diego Division, La Jolla, California.

Converging Sciences (2005), Trento December 2004 Conference, Papers and Transcriptions, University of Trento, Trento, Italy, and Microsoft Research, Cambridge, UK.

Cyranoski, D. (2005), Singapore: An irresistible force. Nature 436: 767-758.

Digitally enabled Genomic Medicine (2004), Fusing Revolutions in Genome-enabled Biomedicine and the Use of the Internet and Information Technology in Society: Projects and People at the University of California, San Diego and Cal-$(IT)^2$, San Diego, California.

Dylan, Bob (1965), "It's alright Ma (I'm only bleeding)" from: Bringing it all back Home, Sony-Columbia Records, New York.

Extending the Internet throughout the Physical World (2001), California Institute for Telecommunications and Information Technology [Cal(IT)^2], California Institutes of Science and Innovation, University of California, San Diego and University of California, Irvine.

Gilbert, W. (1991), "Towards a Paradigm Shift for Biology." Nature 349: 99.

Inventing the Collaborative Research Environment for the Digital Future (2005), A Partnership between UC San Diego and UC Irvine, California Institute for Telecommunications and Information Technology - calit2, La Jolla and Irvine, California.

Lee, D., Lin, A.W., Hutton, T., Shinji, S., Lin, F.P., Peltier, S., Ellisman, M.J. (2003), "Global Telescience featuring IPv6 at iGrid2002." Future Generation of Computer Systems 19: 103139.

National Collaboratories (1993), Applying Information Technology for Scientific Research, National Research Council, NAS Press, Washington, DC.

Peltier, S., Lin, A.W., Mock, S., Lamont, S., Molina, T., Wong, M., Martone, M.E., Ellisman, M.H. (2003), "The Telescience Portal for Advanced Tomographic Applications." Journal of Parallel and Distributed Applications 63: 539-550.

Porter, John et al. (2005), "Wireless Sensor Networks for Ecology." Bioscience 55: 651-572.

PRAGMA (2004), Annual Report 2004-2005, Pacific Rim Applications and Grid Middleware Assembly, Collaboration Overview, University of California, San Diego, Center for Research on BioSystems, San Diego Supercomputer Center, and California Institute for Telecommunications and Information Technology, San Diego, California.

Stokes, D. (1997), Pasteur's Quadrant: Basic Science and Technology Innovation. Brookings Institute Press, Washington DC.

Wooley, John C. (1999), "Trends in Computational Biology: A Summary Based on a RECOMB Plenary Lecture, 1999." Journal of Computational Biology 6: 459-474.

# A Grand Challenge for Converging Sciences

Ronan Sleep

University of East Anglia,
School of Computing Science,
Norwich, NR4 7TJ,
United Kingdom
mrs@cmp.uea.ac.uk

## 1 A Focus for Convergence of Computational and Life Sciences

We routinely use massively powerful computer simulations and visualisations to design aeroplanes, build bridges and to predict weather. With computer power and biological knowledge increasing daily, perhaps we can apply advanced computer simulation techniques to realise computer embodiments of living systems. This is the futuristic proposition of a research challenge proposed by UK computer scientists. The project, called in Vivo – in Silico (iViS) aims to realise fully detailed, accurate and predictive computer embodiments of plants, animals and unicellular organisms.

Initially the aims will be restricted to simple and much studied life-forms such as the nematode worm, the humble weed Arabidopsis, and single cell organisms such as streptomyces and bakers yeast: the worm, the weed and the bug. These model organisms, apparently so different, have much in common:

> *As we trace the increase of complexity from single cell creatures, through small animals like worms and flies . . . evolution is not so much adding new genes performing wholly new functions - what it's chiefly doing is to increase the variety and subtlety of genes that control other genes*[1]

Further, the human and worm have a common ancestor from which we jointly inherit many similar genes. An example is the recent discovery that there is a gene in the worm that is similar to the human breast and ovarian cancer gene BRCA1 (Boulton et al, Current Biology, V14, No.1 pp33-39). So there is considerable hope that studies of the simpler life forms will have real relevance to humans.

Possible benefits of iViS include an understanding of regeneration processes in plants and animals, with potentially dramatic implications for disease and accident victims. But iViS may also lead to revolutionary ways of realising complex systems: instead of designing and programming such systems in excruciating detail, perhaps we can just grow them from compact initial descriptions in a suitable medium. We know it's possible, because that's exactly what nature does with the worm, the weed and the bug.

[1] http://www.sanger.ac.uk/HGP/publication2001/facts.shtml

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 38 – 43, 2006.

## 2 The Vision

iViS offers a powerful vision in which future life scientists can take virtual reality fly-through tours of a plant, animal or colony of cells, studying what is happening at scales ranging from whole life-form to what goes on inside an individual cell, and stretching or shrinking time. Filters control what is seen by the observer, allowing concentration on specific aspects such as cell division, motility or chemical potential.

This is an attractive way of exploring our knowledge about a life-form. But iViS may offer more: with sufficient effort, it might be possible to raise the faithfulness of the underlying model to the point where it becomes predictive as well as descriptive. If this happens, it will become possible to perform meaningful observations and experiments in Silico. And we want to cover a wide range of phenomena: specifically, we include: development from an initial fertilized cell to a full adult, cell function and interaction, motility and sensory behaviour, including interactions with other life-forms. Virtual experiments (e.g. moving a virtual cell during development) should lead to the same outcomes as real life.

### 2.1 iViS and the Life Science Data Mountain

Computers are vital to the Life Sciences: they record, process, visualise and automatically distribute data, and even design and run experiments. They analyze the results in terms of large statistical and other mathematical models of biological processes.

But what do all the numbers, graphs, and spaghetti-like diagrams that emerge from the latest experiments all mean, and what can we do with this data? Making it all fit together into a coherent and useful picture presents a major challenge - many biologists would say the major challenge - facing the life sciences.

This problem is now so pressing that the UK's Biotechnology and Biological Science Research Council (BBSRC) is establishing a number of Centres for Integrative Systems Biology. These Centres will need the vision, breadth of intellectual leadership and research resources to integrate traditionally separate disciplines in a programme of international quality research in quantitative and predictive systems biology. iViS offers a challenging focus of attention for such centres.

### 2.2 iViS as a Driver for Global Knowledge Integration

Part of the answer to the data mountain may lie in the way in which the world wide web is revolutionising our approach to knowledge organisation. The web is already a vital window on the world for scientists wishing to remain up to date. Groups of laboratories that previously worked at arms length and communicated infrequently only via journals and the conference circuit now converse via the web within seconds, swapping massive datasets to compare results. Scientists have begun to exploit the web in its own right by establishing global Virtual Knowledge Repositories to share data, theories and models.

A particularly relevant example is Physiome[2] , which supports

> *the databasing of physiological, pharmacological, and pathological information on humans and other organisms and integration through computational modelling. 'Models' include everything from diagrammatic schema, suggesting relationships among elements composing a system, to fully quantitative, computational models describing the behaviour of physiological systems and an organism's response to environmental change*

Virtual Knowledge Repositories like Physiome will help reduce the proliferation of models and theories that explain parts of the global mountain of life science data. But this in turn will create a deeper challenge: instead of fitting raw data pieces together, we will be faced with the problem of making the models fit into a consistent larger model. Sometimes this will be easy, for example when there is a simple input-output relationship between subsystems. More often — perhaps the rule rather than the exception— combining two models will show unexpected interactions inconsistent with in vivo data.

Part of the problem is that mathematical models deal in numbers, devoid of meaning. The latest evolution of web technology — the semantic web — may help fix this. There is now provision for the web to enhance raw data with additional information called metadata. This can tell the recipient what the data means, how it is represented, the way in which it was generated. Models, which often come in the form of a computer program, can be tagged with metadata describing their assumptions and use: effectively an inbuilt instruction manual.

There are already over 40 metadata dictionaries[3](called ontologies) for the life-sciences. So the drive and energy to create bio-ontologies is already very active. But there is not the same drive to draw them together into a unified whole. The iViS challenge provides just such drive, because the in Silico modelling of a complete life-form, will require harmonious working across all relevant ontology boundaries.

Even if we can build a simulation of a life-form that successfully integrates all known data, we need to take care in choosing our models. If they follow all the raw data too closely, the models may lack any predictive power. For example, we can always find a polynomial of degree (n – 1) to fit n data points exactly. This is a strong reason for complementing data-driven modelling work on iViS with more abstract top-down approaches. If we take care there will be at least some domains which succumb to iViS's whole life form modelling approach: developmental biology looks a good bet.

## 3 Meeting the Challenge: iViS Research Themes

The obvious targets for iViS models are the organisms selected for special attention by biologists for over a century. These range from single cell life-forms such as yeast or streptomyces, through model plants such as Arabidopsis and maize to creatures such as the nematode worm, the fruitfly, and the squid.

[2] http://www.physiome.org/
[3] http://obo.sourceforge.net/

But how can we 'breathe life into data' via computer simulation? This is not simply a question of computing speed or memory, but how to represent the mass of known data as a set of interacting computational processes. We can get computers to simulate massively complex aircraft or bridges, but getting them to grow a worm, weed or bug is significantly beyond the current state of the art.

Nevertheless, it may not be impossible if we build determinedly on the considerable body of work underway to explore ways of organising life science data.

One example is the Edinburgh Mouse Atlas Project4[4]:

> *The EMAP Atlas is a digital Atlas of mouse embryonic development. It is based on the definitive books of mouse embryonic development . . . yet extends these studies by creating a series of interactive three-dimensional computer models of mouse embryos at successive stages of development with defined anatomical domains linked to a stage-by-stage ontology of anatomical names.*

It can be expected that growing numbers of life science virtual knowledge centres will follow EMAP in adopting some form of spatio-temporal framework. The role of iViS is to expand this vision to a dynamic 3-D working model, initially targeting much simpler life-forms.

There are a number of research strands in the Computing Sciences needed to support the aspirations of iViS. We might bundle them under the heading: Computational Models and Scaleable Architectures for in Silico Life Sciences. Some strands will work bottom-up, paying great attention to biological data. Other strands will work top-down, studying minimal abstractions capable of generating the phenomena exhibited in vivo. Many will work 'middle-out', balancing the desire to be simple, elegant and general with the desire to be faithful to the data.

Key to success will be the development of a new breed of computer languages for representing and manipulating biological data in a meaningful way, and using it to drive a realistic, highly detailed, simulation which can be explored using advanced interfaces.

Groups of computer scientists are already exploring new languages, architectures, and system design and analysis tools for the life sciences. Luca Cardelli of Microsoft[5] gives a good picture of this work. Cardelli and others are tackling the complexities of life science head on, developing industrial-quality models aimed at handling the masses of detail in a living system, and - of critical importance if the results of iViS are to be trusted - validating the resulting models.

The impact of such work on the life sciences could be as dramatic as the discovery of structured programming was for computing in the late 1960's. Prior to structured programming, even short programs looked like a mass of spaghetti, just as much of our knowledge of living systems does now. If biological analogues of the compositional primitives of structured programming (sequencing, alternation, and repetition) could be discovered, the prospects for integrated systems biology would be very bright indeed.

---

[4] http://genex.hgu.mrc.ac.uk/

[5] http://www.luca.demon.co.uk/

Such direct attacks on the compexity of biological detail are complemented by more abstract top-down approaches. These begin by asking what sort of computational systems have emergent life-like properties. Such abstract models can be useful when viewing some particular aspect of a plant, animal or cell. For example Prusinkiewicz[6] has almost created a new art form for constructing good-looking pictures of plant growth from remarkably simple abstract models called L-systems. These models capture only a tiny part of the truth, but iViS may need such simplifying ideas to help structure the great mass of detail.

What about raw computer power and advances in haptic and other interface technologies? Certainly they will be needed: but some will emerge anyway from the computer industry without special prompting. The critical problem is to get the underlying computational models right: once we have these, we can begin —as it were— serious work on designing the building and sorting out exactly which of the contemporary technologies we should use to actually construct an iViS prototype.

## 4 Demonstrators and Outline Roadmap

iViS progress will surface as a number of key demonstrators, culminating in whole life form models covering a wide range of phenomena. Intermediate demonstrators will cover a narrower range. Modelling the development of form during development is one example, motivated by the following quote:

> *Perhaps no area of embryology is so poorly understood, yet so fascinating, as how the embryo develops form. Certainly the efforts in understanding gene regulation have occupied embryologists, and it has always been an assumption that once we understand what building blocks are made, we will be able to attack the question of how they are used. Mutations and gene manipulations have given insight into what components are employed for morphogenesis, but surely this is one example where we need to use dynamic imaging to assess how cells behave, and what components are interacting to drive cell movements and shape changes (Scott E. Fraser and Richard M. Harland, Cell, Vol. 100, 4155, January 7, 2000)*

A speculative timeline is:

> **within 5 years:** early results on developmental phenomena in plants and animals, and first unicellular demonstrations. Prototype modelling frameworks and validation methodologies.
> **within 10 years:** first prediction of a textbook result from an assembly of component models; models of meristem growth; models of simple animal development; reliable unicelular models; mature iViS modelling environments.
> **2017**, 100 years after the publication of D'Arcy Thompson's paper 'On Growth and Form', first substantial demonstration of iViS whole life model.
> **within 20 years:** iViS models in common use.

[6] http://www.cpsc.ucalgary.ca/Research/bmv

## 5 What Happens Next?

A website[7] for the iViS Grand Challenge has been established, to help act as a focal point for iViS challenge related work and issues. The narrower the focus, the better the chances of success with the chosen target. On the other hand, too narrow a focus may cause us to miss out vital aspects to whole plant or animal modelling.

In the early stages, it is expected that attention will focus on a number of demonstrators - for example embryogenesis in Arabidopsis. It is hoped that the iViS website will help identify and focus in on a sufficiently small subset of demonstrators with related challenges.

Similarly, the scientific and technical issues identified need to be both broad and deep enough. Those making serious attacks on the challenge will need the mental agility to move attention from one issue to another as each obstacle is overcome.

## Acknowledgements

The general background to the iViS challenge, together with many key references, are available on the iViS website[8]. This website hosts a discussion forum, intended to help mature various aspects of the challenge, and identify a suitably sharp set of scientific questions which need to be addressed if the iViS challenge is to be met. Ioannis Elpidis set up the iViS web site, and helped in preparing this paper.

Special thanks are due to Tony Hoare for his rapid and insightful comments on early drafts of the iViS challenge. He and Robin Milner led a series of UK workshops intended to identify a number of Grand Challenges for Computing Science in the new millenium (see the Grand Challenge website[9]).

The iViS challenge is one of these. It builds on early proposals for a challenge in the interface between computing and biology, most notably: Harel's proposal to model the Nematode Worm, Ross King and Ashwin Srinivasan's proposal to develop a cell model for yeast, and the author's proposal to develop a computational model of gastrulation.

At a more general level, Ray Paton proposed Biology as the challenge for Computing Science in the new millenium; a proposal by Mike Holcombe emphasised the integration issue across a hierarchy of models as critical for credible biological modelling. The complex systems design aspect of the challenge were also present in early proposals: Julian Miller and Catriona Kennedy raised the issue of how to construct systems consisting of huge numbers of interacting elements, and Leslie Smith's proposal focussed on sensory-motor issues.

[7] http://www.cmp.uea.ac.uk/ivis
[8] http://www.cmp.uea.ac.uk/ivis
[9] http://www.nesc.ac.uk/esi/events/Grand Challenges

# Applying Computer Science Research to Biodiversity Informatics: Some Experiences and Lessons

Andrew C. Jones

Cardiff University, School of Computer Science, Queen's Buildings,
5 The Parade, Cardiff CF24 3AA, UK
Andrew.C.Jones@cs.cardiff.ac.uk

**Abstract.** In this paper we discuss experiences of applying Computer Science research in five biodiversity informatics projects. The need that these projects share in common is to apply advanced computing theory and techniques to real problems, overcoming the practical difficulties that are encountered. The main purpose of this paper is to illustrate, from our own experience, how applying advanced computing techniques to a real problem area can lead to unexpected difficulties that may not have been recognised at the theoretical or small-scale implementation stage, and to provide some recommendations for addressing the specific difficulties identified here. These recommendations are in the areas of identifying and rectifying terminological conflicts; handling multiple opinions; and design of architectures that can accommodate variation in platforms and administrative policies, and also accommodate change.

## 1 Introduction

The study of biodiversity is characterised by the need to reason with diverse kinds of information such as species distributions and climate data, and is greatly complicated by the variations between experts concerning the classification of organisms. Biodiversity informatics provides the opportunity to automate and integrate reasoning and computational tasks that would otherwise require substantial manual effort. At the same time, however, the availability of large amounts of biodiversity-related information highlights problems associated with issues such as data quality and variations of terminology used in description and classification. There is a need to capture human expertise and to automate, as far as possible, the resolution of these problems.

In this paper we discuss primarily our own experience of working in specific biodiversity informatics projects, in each of which we have used techniques or software that had not yet been fully stabilised for major applications, and we document our experiences. We recognise that this concentration upon our own work means that the present paper cannot be regarded as a full survey of biodiversity informatics architectures and techniques, but it does draw attention to what we perceive to be some of the most important issues. Some of our solutions to the problems encountered are of general interest, and we outline them here.

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 44–57, 2006.

The remainder of this paper is organised as follows. We commence by providing specific scenarios to illustrate applications of biodiversity informatics. This is followed by a discussion of our experience in the SPICE project (building a federated catalogue of life); GRAB and BiodiversityWorld (building biodiversity e-Science environments); LITCHI (solving problems relating to information retrieval with differing classifications); and the recently-commenced myViews project (supporting multiple opinions and viewpoints within an information system). In the concluding section we summarise the main lessons learned.

## 2 Applications of Biodiversity Informatics

Biodiversity informatics has a wide range of possible applications (for examples see [1,2]). Perhaps an ultimate goal might be to build an e-Science platform in which very disparate data, from molecular data through to information about groups of organisms (e.g. species distribution), as well as other kinds of information such as ecological data, were made available in a sophisticated, exploratory environment. One might envisage an environment in which new hypotheses, of kinds which can only be tested by reasoning with data of such diverse kinds simultaneously, could be expressed and tested. This would involve use of specialist biodiversity resources, but also of resources built primarily for the more mainstream bioinformatics community, and resources from other communities too. This vision relates closely to one of the foci of the Life Sciences Grid (LSG) research group of the Global Grid Forum[1], but its full realisation is still well in the future. Nevertheless, significant progress has been made in certain application areas, and in the remainder of this section we shall describe two of these in order to illustrate the issues that arise: biodiversity richness analysis and bioclimatic modelling (for fuller descriptions of each of these tasks see [3,4]).

In biodiversity richness analysis, a scientist may wish to establish the range of organisms present in a given geographical region. Various metrics have been developed which, given this information, will provide a measure of the biodiversity of the region. Specimen distribution data can be used as a basis for these calculations, but the situation is complicated by a number of issues:

- Scientists disagree about classifications of organisms, and this manifests itself in differences of naming. So, for example, two specimens that belong to the same species may have different names associated with them. It is difficult to ensure that *all* the required species, and *no* others, are retrieved when a scientist is studying a particular group.
- The geographical region may be specified using differing standards and conventions.
- Some of the data may be *wrong*. For example, it is not uncommon for a negative sign to be omitted from a longitude datum, placing the specimen to the East, instead of to the West, of the Prime Meridian.

[1] https://forge.gridforum.org/projects/lsg-rg

Similarly, in the field of bioclimatic modelling, a typical question might be 'Where might a species be expected to occur, under present or predicted climatic conditions?' A question of this sort can be expressed very simply at a high level, but answering it involves using many different kinds of data and tools in order to perform complicated analyses, including a catalogue of life to address the naming problem identified earlier; species information sources containing descriptive or distribution information; geographical data such as climate surfaces; and climate prediction tools such as GARP.[2] A scientist may commence by selecting a particular species or other group of organisms, and building an *ecological niche model* that characterises the regions the species presently occupies. This model can be used for various purposes, including predicting potential invasive species problems and assessing the possible effects of predicted climate change scenarios.

In scenarios such as the two we have just considered, various other problems may arise. For example, originally someone may have created his or her own database just for individual use for some particular purpose, but it may turn out that the data is useful in a more general setting. So the data needs to be made available using appropriate mechanisms, and problems relating to heterogeneity arise. Heterogeneity of various kinds can arise, for example:

- Various representations may have been adopted.
- Granularity of representation may vary. For example, in geographical species distribution data sets, there is variation in the resolution at which the data has recorded.
- In addition to the scientific naming issues identified above, there is the more general problem that an individual user would benefit from seeing other people's data according to his or her own scientific views (e.g. using the user's preferred scientific names); and may wish to selectively include or exclude data according to its provenance.

In the following sections we shall describe our experience in five projects that have aimed to deal with some of the major problems that need to be solved in order to realise an experimental e-Science environment such as we envisage:

- The SPICE project has as its aim the design of a suitable architecture for a distributed, heterogeneous catalogue of life, and the creation of that catalogue. This architecture provides one possible approach to solving interoperability problems, as we shall see. The catalogue provides assistance in solving the problems posed by scientific naming mentioned earlier.
- The GRAB and BiodiversityWorld projects aim to demonstrate the feasibility of developing a Grid [5] for biodiversity research. In the latter project we have developed a general interoperation framework, somewhat different from the interoperation framework developed for SPICE because of the differing requirements, and we have developed tools to discover and orchestrate the use of relevant resources. These projects thus seek to address the need for an experimental environment for biodiversity researchers.

[2] http://www.lifemapper.org/desktopgarp/

- The LITCHI project deals with another more specific problem not directly addressed in GRAB or BiodiversityWorld. It focuses on scientific naming, but a different level from SPICE. It compares *taxonomic checklists*, and tests individual checklists for consistency, identifying conflicts and helping either in their resolution or in mapping between checklists. Thus, the SPICE catalogue can be checked; but also, a more precise selection of scientific name to be used in a given circumstance can be made, as we shall explain.
- The myViews project addresses a related problem, namely that a scientist may wish to sift through and organise large amounts of potentially inconsistent data, imposing his or her own view on it to aid in its interpretation.

## 3 Experiences Gained Implementing Systems to Support Biodiversity Research

### 3.1 Building a Federated Information System: The SPICE Project

The Species 2000 project[3] aims to provide a catalogue of all known species. Due to the distributed nature of the expertise, the fact that individuals normally have expertise in a limited range of species, and the fact that in some cases they have already built databases for their own purposes, it was decided to create a federated architecture in which data providers retain control over their data. A key feature that has simplified the architecture is that Species 2000 specified the data that it required from each of these *global species databases* (GSDs). But it was also necessary to create an architecture that was scalable. The SPICE (SPecies 2000 Interoperability Co-ordination Environment) project had as its main aim an investigation into how a suitable architecture could be created for this distributed catalogue [6]. Figure 1 illustrates the architecture adopted. Key points to note are:

- Given the firm requirements provided by Species 2000 for the data that should be obtained, and the fact that any additional data beyond these requirements did not need to be exposed by individual GSDs, it was possible to define a *common data model*, and to wrap the GSDs so as to conform to this data model. In effect this provided a tightly coupled database federation model. This approach greatly simplified the processing required in the common access system (CAS), and it can be highly recommended in situations where it is feasible.
- We compared two middleware alternatives: CORBA and a CGI/XML-based approach. As anticipated, the CORBA implementation scaled up much better than the CGI/XML-based one; however, practical limitations were encountered in using CORBA outside any individual intranet: in principle a simple HTTP-based approach was found to be much more appropriate despite performance limitations. We discovered that although at the time CORBA seemed an attractive technology to use, in practice it was very difficult to use through firewalls. Also, deploying CORBA on remote sites, where these databases were, led to deployment and configuration difficulties. [7]

[3] http://www.sp2000.org/

– Initially the data providers were reluctant to have their data cached within the CAS. It was demonstrated, however, that a cache-based approach significantly improved performance, as might perhaps have been expected in our particular scenario [7]. This cache is part of the CAS knowledge repository illustrated in our figure.
– Because a very well-prescribed set of operations was required of each GSD (e.g. the ability to provide species names matching a given search string, and to provide individual species data), it was decided to implement these operations, instead of requiring some general query facility to be provided by the wrappers.

It should be borne in mind that the SPICE project commenced when Web services, SOAP, etc., were quite immature; hence the decision not to adopt them at the time. We have now developed SPICE further, supporting a SOAP interface for clients of the CAS so that it can be used as a 'synonymy server'. This means that it can provide information programmatically that may be used to enrich queries of other data sources such as specimen databases, as described earlier. But our findings with respect to HTTP- and CORBA-based implementations illustrate a general point, namely that there is often a trade-off between performance and ease of deployment. In our case the performance issues were made less important by the introduction of caching mechanisms; also, GSD wrapper writers generally preferred to adopt our HTTP-based approach. Again, such caching mechanisms can be particularly effective in situations such as ours where the

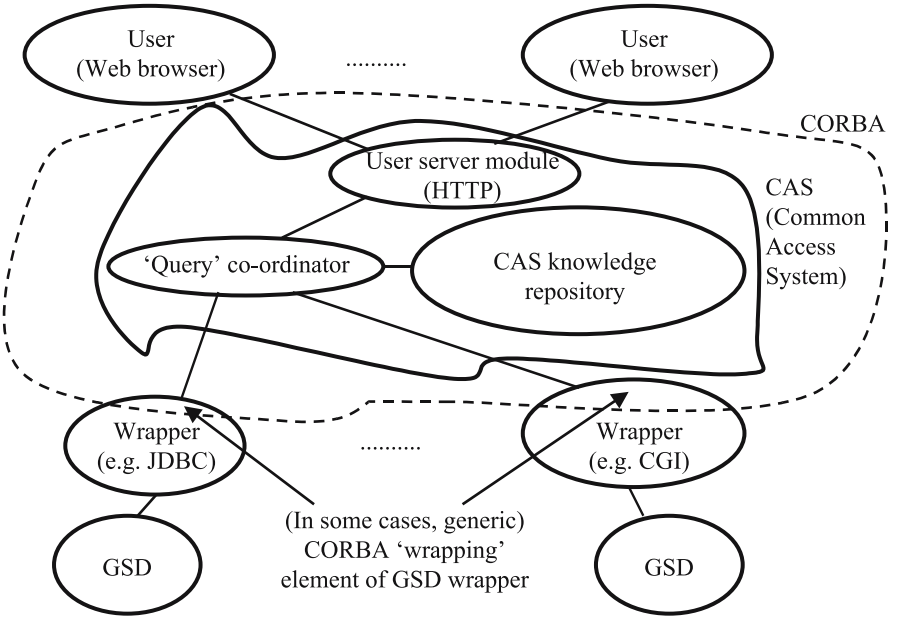



**Fig. 1.** The SPICE architecture

data is changing slowly – often there may be no updates to an individual GSD for several months – and the data is small enough in volume for it to be possible to cache it in its entirety. Similarly, the use of a common data model and the definition of a small set of services required of each component database were made feasible by the nature of the application, and this approach is attractive where it can sensibly be adopted.

### 3.2 Building a Problem-Solving e-Science Environment: The GRAB and BiodiversityWorld Projects

The aim of GRAB (GRid and Biodiversity) [8] and BiodiversityWorld [9] has been to explore how e-Science environments can be designed to support biodiversity research. In both projects we have used the Globus Grid software[4]; indeed, one of the main aims of the GRAB project was to evaluate the appropriateness of Globus for our purposes.

In the GRAB project we built a demonstrator that allowed users to progress through a predefined sequence of operations: retrieve information for selected species including (country-level) geographical distribution; select and retrieve climate data for one of these countries; specify a 'climate envelope', which is initially specified using the country climate data; search for countries having climate within that envelope; retrieve species known to be native to one of these countries. This cycle may be repeated by selecting one of the native species and proceeding as before.

The approach we took to building this software was initially to create a set of services which could be invoked using HTTP requests, and which returned the required data as XML documents for further processing. We discovered that, although Globus (version 2) provided facilities that were useful for certain high-performance computing tasks, it was unnecessarily difficult, at the time, to simply send a request to a remote computer and receive a response, emulating the initial prototype, and that it was difficult to find appropriate documentation to assist us in this task. It was necessary to take an approach involving tasks that created temporary files, etc, which proved a less direct approach to implementing the prototype than our initial HTTP-based approach.

BiodiversityWorld is a follow-on project to GRAB: we are building a much larger-scale problem-solving environment for biodiversity applications. We are investigating how to design a problem-solving environment suitable for the investigation of a wider range of biodiversity problems, and this is being supported by the implementation of a series of prototypes. Of particular importance is the ability to create workflows in a flexible way (we are using the Triana software[5] for this purpose), the ability to incorporate a wide range of resources, and the provision of a sustainable system. Figure 2 illustrates the architecture we have adopted. Of particular note are the metadata repository and the BiodivesityWorld-Grid Interface (BGI). One of the metadata repository's roles is to support resource

[4] http://www.globus.org/
[5] http://www.triana.co.uk/

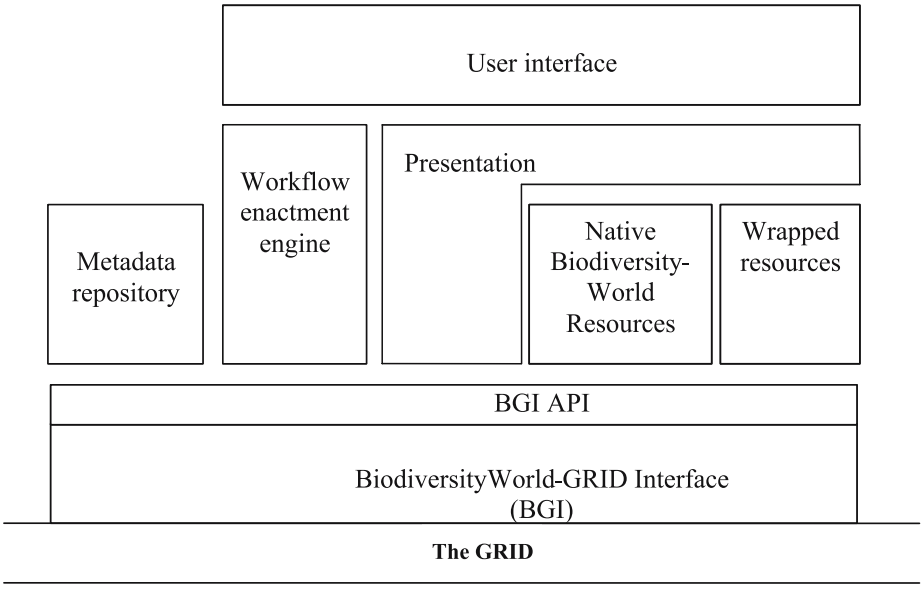



**Fig. 2.** The BiodiversityWorld architecture

discovery: metadata regarding the known resources is important to assist users in selecting appropriate ones for their workflows. Another role is to assist in assembly of workflows: metadata is used to check compatibility of workflow units, and to inform decisions as to any type conversions needed. The motivation for the BGI is that although Globus software has evolved significantly since we used it in GRAB, and it now provides a usable invocation mechanism (Grid Services), it continues to evolve: our concern is that by implementing resources directly as Grid Services they may prove difficult to maintain as the Globus middleware evolves. For this reason we have introduced an abstraction layer between resources (and clients such as our workflow system) and the Grid middleware. This layer provides an invocation mechanism and other features: resources are made available to BiodiversityWorld by wrapping them to conform to the BGI specification, and registering their presence in the metadata repository.

An important issue arising from these projects, as from the SPICE project, is the balancing of performance against other criteria. Clearly our additional middleware layer has an associated performance penalty. But the benefit of this layer is that evolution of resources and of Grid middleware are decoupled, providing us with a more sustainable environment than would otherwise have been created. Another important issue is that there are dangers inherent in committing oneself to early, not fully documented versions of software – particularly middleware – on which the rest of a system depends. Our brief in the GRAB project included the investigation of Globus; otherwise we would have probably delayed using this software until later, more fully featured versions became available.

### 3.3 Information Retrieval and Differing Classifications: The LITCHI Project

In the LITCHI (Logic-Based Integration of Taxonomic Conflicts in Heterogeneous Information sources) project [10] we have been seeking to identify conflicts between species checklists (many of which arise out of differing professional opinion regarding how the organisms concerned should be classified) with the intention of either making the lists consistent or creating mappings between them. In biology there are codes of nomenclature which lead to differences of naming as a result of the differences in classification. An example of the problems with scientific naming is that people use different names in order to refer to the same organism and so in species related data some of the species data will be stored under one name; some under another name. LITCHI has had two phases thus far: in Phase 1 we implemented a stand-alone system requiring data files in a specific format (XDF); in Phase 2 (as part of the Species 2000 Europa project) LITCHI has been reengineered to support the checking of, and mapping between, databases wrapped to be comply with the Species 2000 standards.

In Phase 1 we used constraints expressed in first order logic to represent what constituted a consistent checklist. This is illustrated in figure 3. In this figure, extracts from two checklists are given for the same species, in which the accepted name differs. Also illustrated is a constraint that would detect this conflict.

Some practical problems were encountered in the implementation of these constraints, and the application of repair techniques to restore consistency:

- Many of the constraints we were trying to represent were most naturally expressed by a combination of a 'hard' constraint and a list of individual *exceptions*. For example, there are *conserved names,* which violate the codes of nomenclature – conserved because they have been in long-term, widespread use. We also discovered that scientists preferred to express their knowledge in terms of general rules and rules that were exceptions to the general rules. The approach we took to this latter problem was to re-express these sets of general rules and exception rules as (much more complex) rules that had no exceptions. But the former problem, where individual combinations of data values violated a constraint, was dealt with by noting exceptions to these constraints and replacing the constraints by ones that explicitly accounted for the fact that an exception might have been noted [11].
- Standard repair techniques, such as [12], can fail to identify some of the updates that one might wish to perform in order to resolve a violation. We have discovered that in practice, an individual constraint may need to be replaced by several constraints, each with additional information relating to the cause of the violation, so that appropriate repairs may be selected [13].
- More generally, a single taxonomic cause (e.g. one scientist regards a group of species as belonging to a single genus; another might regard them as divided into two genera) may be manifested in potentially very many conflicts between individual names in a pair of checklists. Our constraints are thus identifying symptoms rather than causes, which has limited us in the

**Extract from list 1:**
Caragana arborescens Lam. (accepted name)
Caragana sibirica Medicus (synonym)

**Extract from List 2:**
Caragana sibirica Medicus (accepted name)
Caragana arborescens Lam. (synonym)

**Constraint expressed informally in English:**

A full name which is not a pro parte name[a] may not appear as both an accepted name and a synonym in the same checklist.

**Constraint expressed in first-order logic:**

$$
\begin{aligned}
&(\forall n, a, l, c_1, c_2, t_1, t_2)( \\
&\quad accepted\_name(n, a, c_1, l, t_1) \\
&\quad \wedge synonym(n, a, c_2, l, t_2) \\
&\quad \Rightarrow pro\_parte(c_1) \wedge pro\_parte(c_2))
\end{aligned}
$$

[a] Pro-parte names arise when it is decided that a taxon must be split into smaller taxa at the same taxonomic rank, e.g. a species is split into two or more new species.

**Fig. 3.** Conflicting extracts from two checklists, and constraint to detect such conflicts

facilities that we have been able to provide for dealing with conflicts efficiently. We were applying integrity repair theory to a real situation where the theory's assumptions do not fully apply. In our case it is not merely an individual transaction that will have led to an inconsistency in a database; rather, our database will have a potentially very large number of inconsistencies, many of which exist because they reflect a much smaller number of differing professional opinions.

- A more mundane problem relates to the software and architecture chosen to implement our system. We chose the Prolog language for the constraint-based part of our software, due to the ease with which constraints could be expressed; Visual Basic for the User Interface, and Microsoft Access for storing the data. We used third-party, not fully supported software to interface between those components, but unfortunately we found that our system only worked on a small subset of the computers that we had access to, for reasons that we were never able fully to isolate.

As a result of these considerations, in Phase 2 we have experimented with a different approach: detecting patterns of relationships between names and inferring taxonomic causes from these relationships. This software is also now implemented entirely in Java, for portability. In this latter phase we concentrate not only upon detecting and resolving conflicts between checklists, but also upon generating cross-maps between the taxonomic views they represent. This means that, if knowledge is held about the taxonomic view underlying a given database

(e.g. a database of specimen distributions) and about the user's taxonomic viewpoint, the correct name(s) for use in searching the database can be selected by using the information held in the cross-map. Experimentation with our new approach is still proceeding, so it is too early to report on the effectiveness of this idea. But lessons that arise from Phase 1 are the need to adapt theory when applied to a real problem (and, in particular, to think carefully about how to express constraints most efficiently); and the dangers associated with building a large research project in dependence upon relatively unstable tools and middleware. In relation to scientists' preference for general rules with exceptions, it would be interesting to explore whether default logics could be employed to capture their knowledge more directly, but we have not yet considered this issue fully.

### 3.4 Towards a System Supporting Multiple Opinions

The LITCHI project deals with one aspect of a more general problem within biodiversity informatics – and, indeed, a problem that in one form or other manifests itself in a wide range of disciplines. This is the problem of divergent professional opinion and practice, which can manifest itself in disagreements over classification; disagreements over terminology; differences of preference about what features of an entity should be described, and variation in the level of confidence that one particular professional might place in another's judgement, to name only a few areas.

In this paper we have seen how scientists may disagree about the namings of organisms. A particularly challenging manifestation of this problem is in situations where the boundaries between different groups of organisms and the number of groups (such as species) that actually exist are in dispute. This can mean that whether a scientist wishing to retrieve data regarding a given species should be provided with data from a given specimen depends both on correctly identifying the classification scheme used to originally label the specimen and on the scientist's own opinion. A similar problem can arise at a descriptive level: in one scientist's opinion the flowers of a particular plant might be "bright yellow"; another might use the expression "golden yellow". Or one scientist might record leaf length as an important measure; another might prefer to record leaf surface area. It is desirable to organise all this information into a form that is useful to an individual scientist, allowing him or her to issue queries that are expressed in his or her own terms, but querying data sources using their own native terms and transforming the results back into a usable form. It is also potentially useful for a scientist to be able to specify which data sets (s)he accepts and which (s)he rejects, based (for example) on his or her opinion of the professional judgement of the originator of the information. The ability to retain data in its original form is also important, so that multiple views can be created above the data depending on the user's current preferences, and multiple users can access the same base data without interfering with each other.

The myViews project [14] is addressing precisely this problem. At present myViews consists of an early prototype illustrating some of the reasoning and

knowledge a system must be capable of if it is to provide the flexible views described above. It is implemented in Prolog which, as a declarative language with built-in inference capabilities, is well suited to the creation of a proof-of-concept prototype. The prototype currently holds knowledge on a small range of *Cytisus* species, taken from various sources.

Each item of knowledge is stored as an *assertion,* which includes:

- information on the source of the assertion, and
- the assertion itself, which is a predicate applied to one or more objects in the system.

Assertions can represent information coming from other scientists, such as descriptions of specimens, descriptions of species, names that are regarded as synonyms, etc, but they can also represent information from individual users. The user may assert, for example, that any assertion about one species is equally an assertion about another species, because the two are synonymous; or that (s)he accepts or rejects assertions from a given source. Examples of such assertions (expressed informally in English) are:

- In publication X, *Cytisus scoparius* has the common name 'Scotch Broom'.
- In publication Y, *Sarothamnus scoparius* is said to have bright yellow flowers.
- In publication Z, *Cytisus scoparius* is said to have golden yellow flowers.
- The user regards any assertion about a *Sarothamnus* species as being also an assertion about the corresponding *Cytisus* species (e.g. assertions about *Sarothamnus scoparius* are also assertions about *Cytisus scoparius*).

In addition, some variation in description may occur which appears contradictory, but the user may have knowledge that it is not a real contradiction, perhaps because scientists had considered different specimens of the same species, but with differing features. For example,

- In publication X the flowers of *Cytisus scoparius* are said to have no scent.
- In publication Z the flowers of *Cytisus scoparius* are said to *have* scent.

The user can assert that such variations are regarded as not being real discrepancies – perhaps in some cases the flowers have scent, and in others they do not. Similarly, when differences of descriptive terms or units arise, rules are encoded which deduce whether these differences are contradictions or not. For example, if the units of one measurement are centimetres and another, inches, a conversion is performed in order to determine whether the measurements are actually equivalent.

The system can thus present the user with the data that (s)he has chosen to accept. It can also infer and present the user with contradictions within this accepted data.

At present myViews is in its very early stages, and no funding has yet been sought in order to support a full-scale exploration of the underlying ideas. A fuller prototype would require development of a suitable user interface; caching

of views created within the system; an interface to persistent knowledge bases; and attention given to scalability issues, using ideas from the deductive database research community and elsewhere to ensure that a myViews system of non-trivial size will be able to work efficiently.

The main significance that should be seen in the myViews project is that although environments such as BiodiversityWorld can assist researchers in discovering knowledge and co-ordinating complex analytic tasks, further support is needed to assist them in making sense of large amounts of data that reflect a variety of scientific opinion.

## 4 Conclusions

In this paper we have shown that biodiversity informatics is a multifaceted research area, and that novel techniques and tools are needed to support biodiversity scientists in their research. We have observed that:

- Dependence on third-party prototype software as a basis for a research project can lead to difficulties.
- Architectures should not be needlessly complicated: for example, in SPICE, full advantage has been taken of the limited range of data that is required to be exposed by component databases.
- There are often trade-offs that need to be considered very carefully between factors such as performance, scalability, ease of deployment and maintainability as parts of a system evolve.
- Surprising problems can arise when one applies theoretical knowledge to realistic problems, as we have done in LITCHI. A corollary of this is that applied computer science can involve tackling challenging research problems associated with the practical implementation of computing theory.

A key area in which further research is needed is into semantic issues surrounding biodiversity data. We have illustrated, in LITCHI and myViews, techniques that can assist in managing the complex semantics associated with data in this field, and continued research is needed to improve the accuracy with which information can be retrieved and used in biodiversity research.

## Acknowledgements

SPICE and LITCHI were originally funded by grants from the Bioinformatics Committee of the UK BBSRC/EPSRC. The second phase of these two projects is currently funded as part of an EU Framework 5 grant. BiodiversityWorld is funded by a research grant from the UK BBSRC research council. GRAB was funded by a grant from the UK Department of Trade and Industry (DTI).

Collaborators in the above projects at the Universities of Cardiff and Reading; the Natural History Museum, London, and Royal Botanic Gardens, Kew,

are gratefully acknowledged, as are a wide range of institutions whose data has been used in these projects.

This is a revised, substantially expanded version of a paper presented at the Microsoft Converging Sciences Conference, December 2004.

## References

1. Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T.: New developments in museum-based informatics and applications in biodiversity analysis. TRENDS in Ecology and Evoloution **19** (2004) 497–503
2. Bisby, F.A.: The quiet revolution: Biodiversity informatics and the internet. Science **289** (2004) 2309–2312
3. Loiselle, B.A., Howell, C.A., Graham, C.H., Goerck, J.M., Brooks, T., Smith, K.G., Williams, P.H.: Avoiding pitfalls of using species distribution models in conservation planning. Conservation Biology **17** (2003) 1591–1600
4. Peterson, A.T., Vieglais, D.A.: Predicting species invasions using ecological niche modeling: New approaches from bioinformatics attack a pressing problem. BioScience **51** (2001) 363–371
5. Foster, I., Kesselman, C., eds.: The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, San Francisco, CA (1999)
6. Jones, A.C., Xu, X., Pittas, N., Gray, W.A., Fiddian, N.J., White, R.J., Robinson, J.S., Bisby, F.A., Brandt, S.M.: SPICE: a flexible architecture for integrating autonomous databases to comprise a distributed catalogue of life. In Ibrahim, M., Küng, J., Revell, N., eds.: 11th International Conference on Database and Expert Systems Applications (LNCS 1873), Springer Verlag (2000) 981–992
7. Xu, X., Jones, A.C., Pittas, N., Gray, W.A., Fiddian, N.J., White, R.J., Robinson, J.S., Bisby, F.A., Brandt, S.M.: Experiences with a hybrid implementation of a globally distributed federated database system. In Wang, X.S., Yu, G., Lu, H., eds.: 2nd International Conference on Web-Age Information Management (LNCS 2118), Springer Verlag (2001) 212–222
8. Jones, A.C., Gray, W.A., Giddy, J.P., Fiddian, N.J.: Linking heterogeneous biodiversity information systems on the GRID: the GRAB demonstrator. Computing and Informatics **21** (2002) 383–398
9. Jones, A.C., White, R.J., Gray, W.A., Bisby, F.A., Caithness, N., Pittas, N., Xu, X., Sutton, T., Fiddian, N.J., Culham, A., Scoble, M., Williams, P., Bromley, O., Brewer, P., Yesson, C., Bhagwat, S.: Building a biodiversity grid. In Konagaya, A., Satou, K., eds.: Grid Computing in Life Science: First International Workshop on Life Science Grid, Revised selected and invited papers (LNCS/LNBI 3370), Springer Verlag (2005) 140–151
10. Embury, S., Jones, A., Sutherland, I., Gray, W., White, R., Robinson, J., Bisby, F., Brandt, S.: Conflict detection for integration of taxonomic data sources. In Oszoyoglu, M., ed.: 11th International Conference on Scientific and Statistical Databases, IEEE Computer Society Press (1999) 204–213
11. Jones, A.C., Sutherland, I., Embury, S.M., Gray, W.A., White, R.J., Robinson, J.S., Bisby, F.A., Brandt, S.M.: Techniques for effective integration, maintenance and evolution of species databases. In Günther, O., Lenz, H.J., eds.: 12th International Conference on Scientific and Statistical Databases, IEEE Computer Society Press (2000) 3–13

12. Wüthrich, B.: On updates and inconsistency repairing in knowledge bases. In: Proceedings of the Ninth International Conference on Data Engineering (ICDE), IEEE Computer Society Press (1993) 608–615
13. Embury, S.M., Brandt, S.M., Robinson, J.S., Sutherland, I., Bisby, F.A., Gray, W.A., Jones, A.C., White, R.J.: Adapting integrity enforcement techniques for data reconciliation. Information Systems **26** (2001) 657–689
14. Jones, A.C.: Some thoughts about computing techniques to support generation of coherent taxon descriptions. In: Proc. 5th Biennial Meeting of the Systematics Association. (2005) http://www.systass.org/biennial2005/index.html.

# e-Science and the VL-e Approach

L.O. (Bob) Hertzberger

Computer Architecture and Parallel Systems Group,
Department of Computer Science,
Universiteit van Amsterdam
bob@science.uva.nl

## 1 Introduction

Now that the Internet and Web are abundantly present in every days live and networking and computing speed are improving at an impressive rate, more and more attempts become visible to apply these developments towards science, in general often called: e-Science.

In order for e-Science to work, an environment has to be created that should support the requirements of different scientific domains such as life sciences or high energy physics research.

Looking in nowadays experimental science one of the most important observations is that experiments become increasingly more complex driven by the increase in detector resolution and automation as well as automation through robots.

For instance, in biology the study of a particular genome is replaced by studying a large amount of genomes in (what is called the genomics) experiments made possible

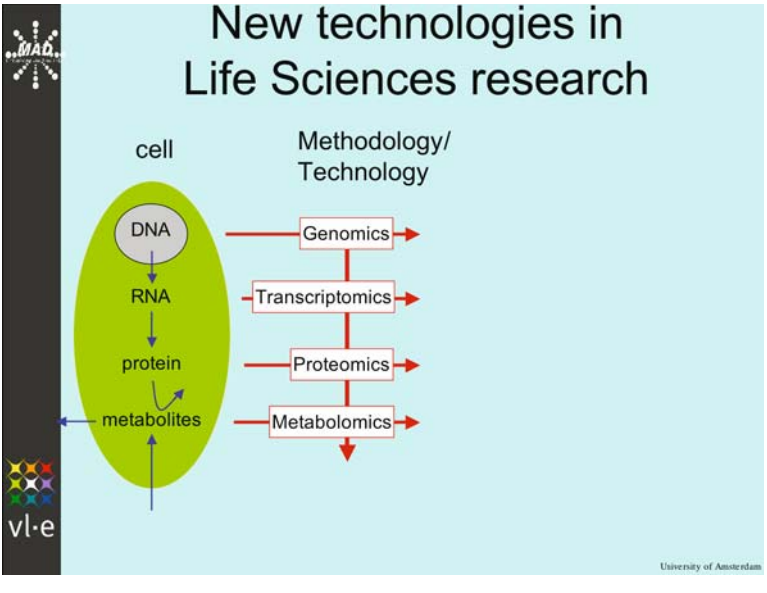



**Fig. 1.** Omics methods in life sciences

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 58 – 67, 2006.

via micro-array equipment. Such high throughput experiments produce large quantities of data that have to be processed and converted into useful life science information. An illustration of this so-called "-omics" influence in life sciences is illustrated in Figure 1.

It is clear that such developments must have an impact on the way one is looking into life science experiments. Past experiments where hypothesis driven to evaluate existing hypothesis with the purpose to complement this knowledge.

Present experiments are data driven in order to discover knowledge from large amounts of data. This inevitably leads to a paradigm shift in life science. In such new experiments statistical techniques are applied to discover new insides from the large data sets being produced.

Moreover data sets of interest are not limited to one's own laboratory but they are distributed all over the world. There for it becomes necessary to have the methodology as well as the permission to access parts of other people's data sets of interest, in order to complement ones own knowledge. That is quite a different approach and request, beside other methods and techniques, another type of ICT based infrastructure.

The result of the increased complexity of experiments is an increase in amount and complexity of the data they produce. This is often called the application data crisis. Some of the consequences are illustrated in table 1:

**Table 1.** Some examples of the application data crisis

| | |
|---|---|
| medical imaging (fMRI): | ~ **1 GByte** per measurement (day) |
| Bio-informatics queries: | ~ **500 GByte** per database |
| Satellite world imagery: | ~ **5 TByte**/year |
| Current particle physics: | ~ **1 PByte** per year |
| Future particle physics): | ~ **10-30 PByte** per year |

Another problem complicating this situation is the fact that data is often very distributed.

In a field like High Energy Physics the impact of new instrumentation was dramatic. In early experiments a particle beam was shot on a target and the resulting particles were studied via photos taken from a so-called bubble chamber. With this set-up only low density experiments could be performed. This situation changed when new instrumental techniques made so-called counter experiments possible. However, these experiments produced a considerable larger amount of data. It lasted a number of years before the big data sets resulting from these experiments could be processed in such a way that the scientifically important information could be extracted. In addition, these instrumental methods resulted into rethinking the experiments in general and produced a completely new type of very high density, the so-called

colliding beam experiments. These developments had an impressive impact on the elementary particle field itself and produced numerous new insights.

It is not suggested here that a completely equal development in life sciences can be expected, but similar developments certainly can not be ruled out. In that respect it could be argued that a conference with as subject converging sciences becomes a bit too early as the domain is not well established yet.

It can be expected that understanding the potential of high throughput experimentation will undoubtedly require a rethinking of experiments themselves. This is often called design for experimentation.

One step forward is to make certain that the heterogeneous data repositories resulting from these experiments can be made more easily available in a networked computer environment. This means that the experimental resources have to be made virtual. In such a way it will become possible to easily create data repositories and retrieve the relevant information whenever and wherever necessary. By doing so an important step towards collaborative experiments and e-Science can be made.

## 2 What Is e-Science?

Web is about exchanging information interpretable by humans. Another development, "Grid", is aiming towards harnessing computer resources and by virtualization making them available for sharing. For science this implies sharing of computing power, data and information repositories, as well as expensive experimental facilities.

More than only coping with the data explosions, e-Science is targeting towards enabling multi-disciplinary science combining human expertise and knowledge between the domain scientist (such as a biologist) and the ICT scientist.

Because the computer is becoming an integrated part of the experiment it will demand a different approach towards experimentation. This will give an additional push towards a radical change in design for experimentation.

In addition e-Science has the possibility to further push the current state of multi-disciplinary research in science forward, because of its potential for collaborative cooperation.

It has to be stressed, however, that this is not only a matter of ICT infrastructure, but more a set of mind. The experience is that the problem for true collaboration often is either scientist of different disciplines discussing what they believe to be the same phenomena, but using completely different terminology or using the same terminology but meaning something completely different.

## 3 What Should Be e-Science objectives?

For e-Science to stimulate the scientific process and consequently being effective the following requirements could be defined:

- Stimulating collaboration by sharing data and information,
- Improve re-use of data and information,

- Allowing for combing data and information from different modalities using sensor data and information fusion,
- Realize the combination of real life and (model based) simulation experiments.

An example of the important role sharing plays for collaborative research can be found in the case of using of fMRI for experimental cognition research. A long debate about the usefulness of data sharing using large repositories of data coming from cognition experiments was recently, at least temporarily, brought to an end by a Nature article [1] showing the potential for the Neurosciences of such undertakings. Based on an existing example it was outlined that a database for sharing studies of human cognition was realised and the advantages such as collaborative research, information sharing, independent experiment validation, training, etc. were discussed.

To make the shared usage of data more successful, it is important to standardize protocols which inevitably is leading to more de-facto rationalization of the experimental process. This has the advantage that the experiment becomes more reproducible and comparable allowing better re-use of experimental results.

The combination of sensor detection modalities is very important. One could think of combining data coming from transcriptomics (micro-array) experiments with that of proteomics (mass spectroscopy) research. This can for instance be used in the case of studying phenomena causing breast cancer. Doing so, it becomes important to have good possibilities to calibrate the datasets.

The potential of combining real life with model based simulation experiments can only fully be exploited when it is realized that it becomes essential for the computational experiment that it can be validated. An example is presented in section 5.

e-Science should result in computer aided support for rapid prototyping of ideas and such stimulate the creativity process. E-Science should be realized by creating and applying new methodologies and an ICT infrastructure stimulating this. One of the important issues is that the results of experiments done with this infrastructure can be rapidly back propagated into the ICT infrastructure. Along these lines we have started our Virtual Laboratory for e-Science (VL-e) project [2].

## 4 Virtual Lab for e-Science(VL-e) Research Philosophy

The aim of the VL-e project is to do multidisciplinary research in different application domains, in order to develop generic methodologies, as well as to develop the necessary ICT infrastructure.

As a consequence the various application cases are the drivers for computer and computational science and engineering research. Bioinformatics will be used here to further illustrate this point.

On the one hand bioinformatics has the scientific responsibility to develop the underlying computational concepts and models to convert complex biological data into useful biological and chemical knowledge. On the other hand it has the technological responsibility to manage and integrate large amounts of heterogeneous data sources from high throughput experimentation such as micro array and mass spectroscopy experiments.

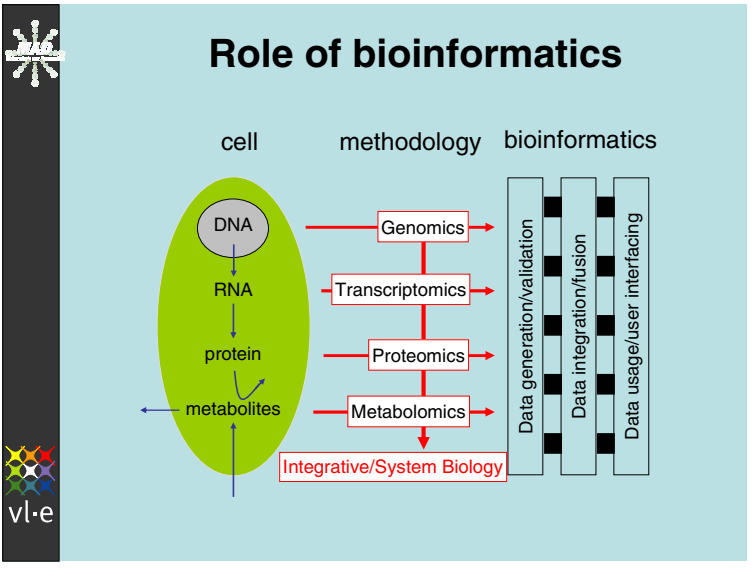



**Fig. 2.** The role of bioinformatics

Bioinformatics is believed to be one of the most important enabling technologies necessary to develop new directions in biology, medicine and pharmacy.

In Figure 2, which is an extension of Figure 1, the role of bioinformatics is further illustrated. In particular the need to handle large heterogeneous databases is critically dependent on the quality of the ICT infrastructure available and illustrates the intertwining of bioinformatics and its infrastructure.

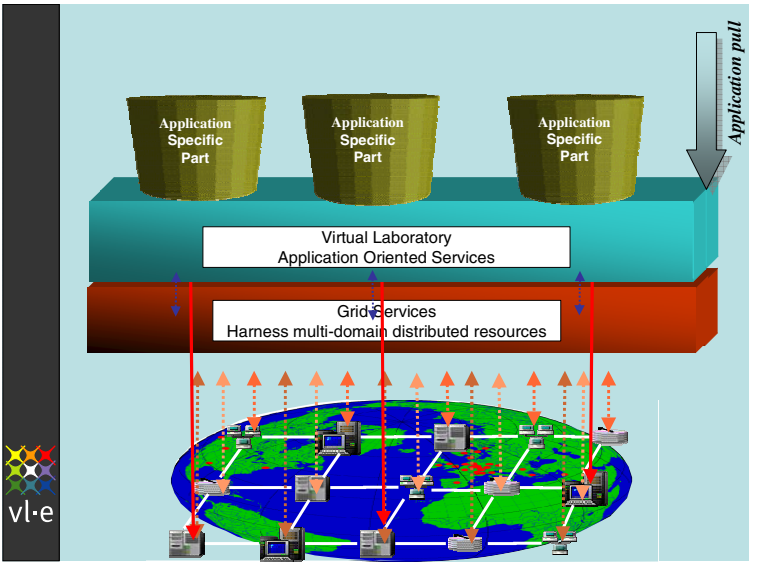



**Fig. 3.** Problem solving environment

To realize application driven research for the various domains in VL-e it was concluded that it was necessary to develop a separate problem solving environment for each of the domains under study (see Figure 3). When studying the commonalities between problem solving environments for different application cases, the conclusion was that essentially two different parts could be distinguished.

One part is application specific for a certain domain. The other, in which such issues as visualization, information management, workflow concepts, aspects in modelling and simulation, etc. are considered, is common for all and consequently generic in nature. Methods and software realizing them can be re-used and can be made part of the virtual laboratory environment (see Figure 3). Consequently the VL-e project extensively uses the concept of problem solving which has a generic and application specific part.

In Figure 4 the project and its various application cases are illustrated. The medical case mainly looks at fMRI applications for medical and cognition research**.** The bio-diversity application is targeting towards studying the problem of integrating large amounts of small databases containing information about species.

The so-called Dutch Telescience laboratory is specifically looking into the usage of large apparatus like mass spectrometers for proteomics and the way the pre-processing of the data has to be handled. Moreover its aim is building data repositories that can be shared. The role of bioinformatics was discussed already. The Food informatics case aims to develop an environment were food ontologies and their role in better designing food processing can be studied.

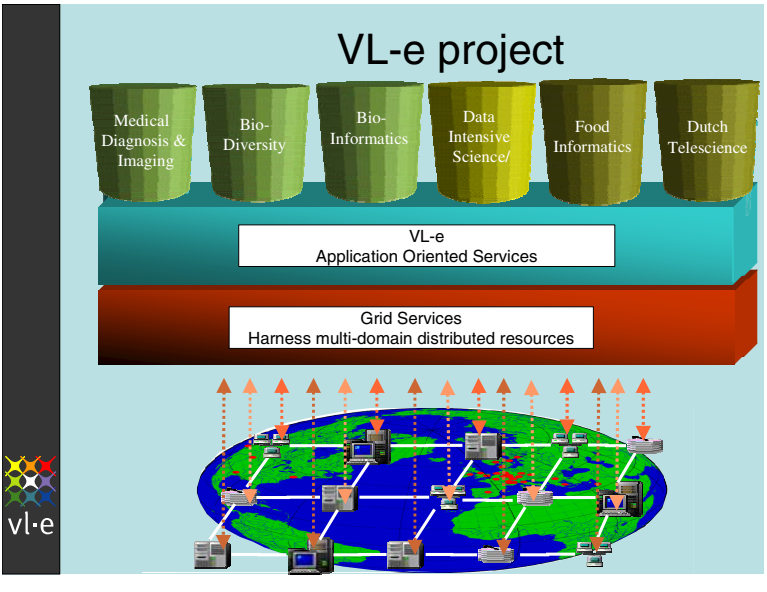



**Fig. 4.** Different VL-e application domains

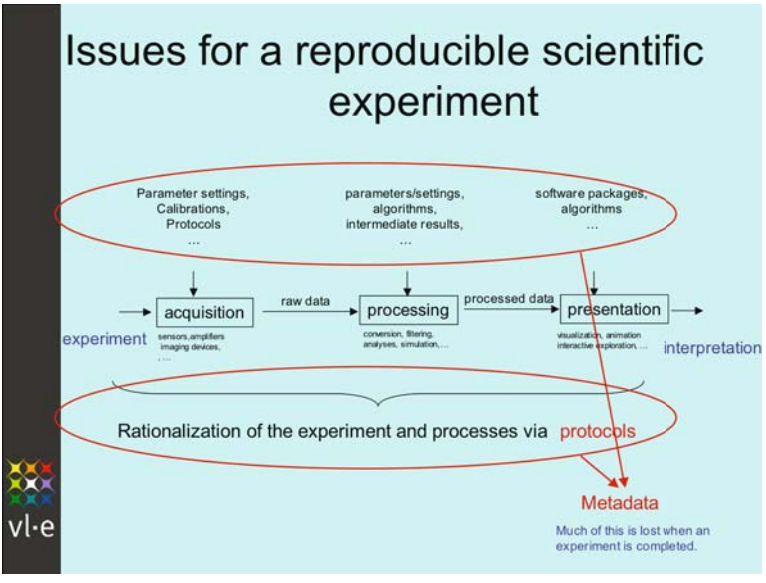



**Fig. 5.** Reproducible experiment

The data intensive science case studies the role of data handling and processing for experiments that are well known for the fact that they produce large quantities of data like high energy particle physics and astronomy.

The fact that the VL-e project intends to re-use software components where and whenever possible can have an impact on the design of the experiment [3]. It was observed already that design for experimentation is important when dealing with new types of instrumentation. The usage of computer methodology as an integrated part of the experimental system can, in that respect, also be considered as a form of experimentation.

This is specifically the case when combining model based simulation and real life experimentation, as well as combining the data coming from these experiments (see section 5).

Especially when collaborative research is important it inevitably has an impact on the way an experiment is undertaken.

All factors discussed here leads to the necessity to further rationalize (and de-facto standardize) the various steps undertaken when carrying out an experiment and consequently make it better reproducible and comparable.

In Figure 5 the issues when designing an experiment in a virtual environment are illustrated. In this simple drawing it is outlined that in all steps of the experimental process information is lost which is playing an essential role when studying the results. To better capture that information it becomes necessary to formalize it in the form of protocols and capture it, among others, via workflows and metadata. This has certain disadvantages because it might also limit the experimentalist in his creativity process. On the other hand, especially when the computer is an integrated part of the experiment, it is necessary to explicitly describe steps in the process in order to improve reproducibility.

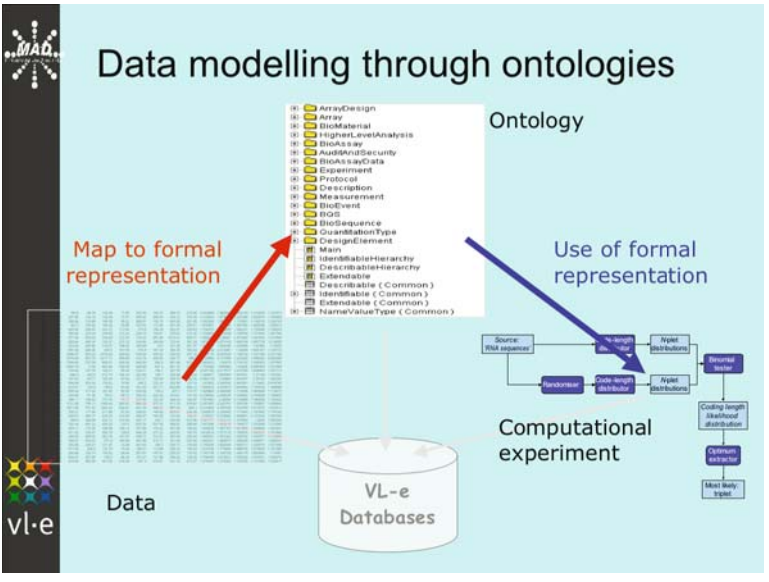



**Fig. 6.** Use of ontologies for data modelling in bio-informatics

The protocols described can be based on workflow definitions to recreate complex experiments into process flows. Such workflows can then play a role in the control over and in the execution of a computer experiment. These can also be used as a basis to configure the various processing modules that are necessary for data processing interpretation and visualization. To define all these activities, ontologies can help to obtain well-structured definitions of experiment data and meta data. In Figure 6 an example of data modelling using ontologies in bio-informatics is presented.

In the VL-e project it was discovered that the combination of application driven ICT research as well as basic computer science research required different computer based environments. The reason being that the application oriented research demanded an environment with a certain amount of stability that would severely limit the creativity as well as the needs of the computer science researchers. A researcher in bio-informatics for transcriptomics should be able to use the infrastructure and concentrate on his topic being developing new software for micro array experiments analysis. This cannot happen if there is not a certain level of stability in the underlying hardware and software infrastructure.

On the other hand an environment is necessary where computer-, computational and domain specific ICT scientist can do their experiments. In addition new ideas in application driven virtual laboratory research might have an impact on, or might demand new facilities from the underlying ICT infrastructure. For this testing a Rapid Prototyping environment was created. Therefore a distinction is made between the latter and a more stable, so-called Proof of Concept environment (see Figure 7).

One of the consequences for making this choice is the problem that software developed in the Rapid Prototyping environment has to migrate towards the Proof of Concept environment and consequently an authentication authority is required that is doing thorough testing before software is migrated.

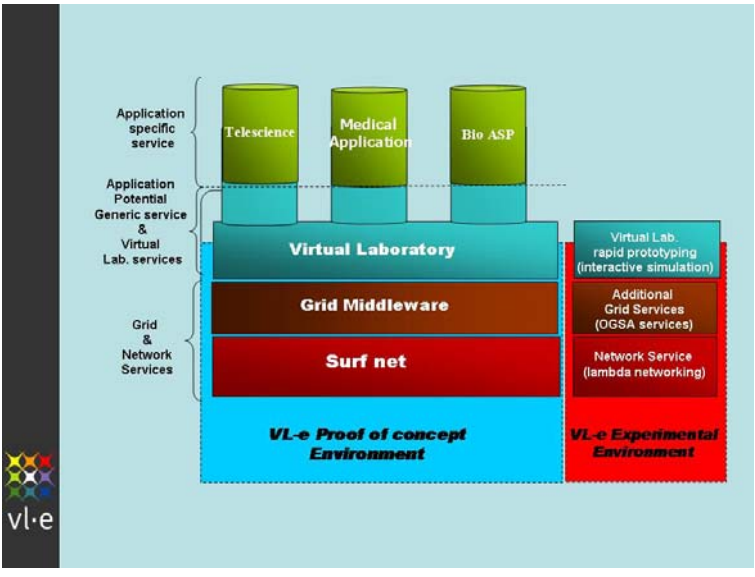



**Fig. 7.** Proof of Concept and Rapid prototyping environment

## 5 e-Science Example

An example of combining (model based) simulation and real life experiments is the problem solving environment for Simulated Vascular Reconstruction [4]. Here the

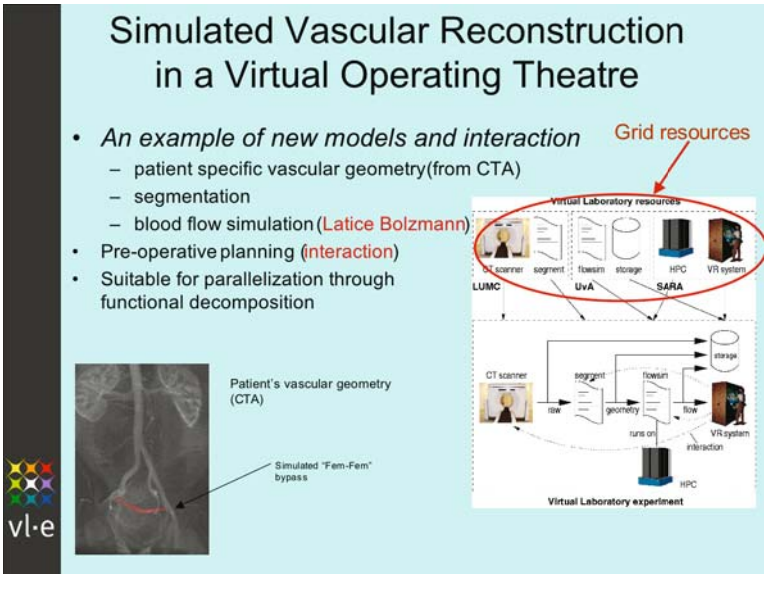



**Fig. 8.** Medical problem solving

aim is to realize computer aided design system for a bypass operation in a virtual operating theatre. It is a good example of model based interaction and computational steering. A patient is put in a CT scanner to obtain the geometry of the blood vessels under consideration. Thereafter a simulation of the bloodstream through the vessels is obtained using Lattice Bolzmann methods. The CT scanner is located in Leiden, whereas the blood flow simulation is carried out on a grid system in Amsterdam.

Visualization and interactive steering is realized in Amsterdam as well as in Leiden. This is an example of an application that has been realize in a distributed environment and be carried out by different groups in the Netherlands working on the same problem but each on a different aspect. It is therefore also a nice example of how to realize multidisciplinary research.

## 6 Conclusion

e-Science is a lot more then coping with the data explosion alone. Especially in the Life Sciences field the so-called "-omics" oriented experiments will result in a paradigm shift.

It was illustrated that future e-Science implementation requires further rationalization and standardization of the experimentation processes. In this activity ontologies might be of help. It was further illustrated that e-Science success requires the realization of an environment allowing application driven experimentation and rapid dissemination of and feed back of these new methods.

Doing so it is not enough to only provide the network, or provide grid middleware for users, let alone let user communities solve their own problems. One has to integrate all this in order to create a multidisciplinary environment where all these factors are coming together.

## Acknowledgements

This work was carried out in the context of the Virtual Laboratory for e-Science project (www.vl-e.nl). This project is supported by a BSIK grant from the Dutch Ministry of Education, Culture and Science (OC&W) and is part of the ICT innovation program of the Ministry of Economic Affairs (EZ).

## References

1. John Darrell Van Hove, Scott T Grafton, David Rockmore & Michael S Gazzaniga, *Nature Neuroscience*, Volume 7, 471-478, 2004.
2. Virtual Lab e-Science, Towards a new Science Paradigm, A Bsik proposal, Feb2003.(www.vl-e.nl).
3. A.S.Z. Belloum, D.J.Groep, Z.W.Hendrikse, L.O.Hertzberger, V. Korkhov, C.T.A.M. de Laat and D.Vasunin. VLAM-G: A Grid-based virtual laboratory. Future *Generation Computer Systems*, 19(2):209-217, 2003.
4. R.G.Belleman and R.Sulakov. High performance distributed simulation for interactivevascular reconstruction. *In P.M.A. Sloot;C.J.K.Tan;J.J.Dongarra and A.G.Hoekstra, editors, Computational Science- ICCS 2002, Proceedings Part III, in lecture Notes in Computer Science, vol2331*,265-274. Springer Verlag, 2002.

# From Syntax to Semantics in Systems Biology Towards Automated Reasoning Tools

François Fages

Projet Contraintes, INRIA Rocquencourt,
BP105, Le Chesnay Cedex 78153, France
Francois.Fages@inria.fr
http://contraintes.inria.fr

Mathematical biology has for a long time investigated the dynamics of biomolecular systems by developing numerical models involving (highly non-linear) differential equations and using tools such as Bifurcation Theory for estimating parameters [1]. Mathematical biology provides a firm ground for the numerical analysis of biological systems. However, state-of-the-art quantitative models can hardly be re-used and composed with other models in a systematic fashion, and are limited to a few tenths of variables [2].

Qualitative models of bio-molecular interactions constitute the core of nowadays cell systems biology. Interaction diagrams are the first tool used by biologists to reason about complex systems. The accumulation of knowledge on gene interaction and pathways is currently entered in databases such as KEGG[3], EcoCyc [4], etc. in the form of annotated diagrams. Tools such as BioSpice, Gepasi, GON, E-cell, etc. have been developed for making simulations based on these databases when numerical data is present. Furthermore the interoperability between databases and simulation tools is now possible with standard exchange formats such as the Systems Biology Markup Language SBML [5].

These advances give more acuity to at least three challenges for systems biology:

– One big challenge is the modularity and compositionality of biological models. It is not an easy task today to combine given models of different pathways sharing some molecular components in a given organism, and obtain a mixed model of the complex system. This is a restriction to the re-use of models in systems biology and to their direct use in any application.
– Another challenge is to go beyond simulations and use models to automate various forms of biological reasoning, in purely qualitative models too. Computer aided inference of interaction networks, or computer aided drug target discovery, need non-trivial automated reasoning tools to assist the biologist.
– A third challenge for systems biology will be the possibility to change the way molecular cell biology is taught by making it more formal, putting formal models and tools at the center of the courses. Having a common syntax, one way to approach these challenges is to develop precise semantics of interaction diagrams and build formal methods and tools to reason about them. Our project with the Biochemical Abstract Machine[1] [6], started in 2002, is such an attempt. Based on formal semantics of molecular interactions, Biocham offers:

[1] http://contraintes.inria.fr/BIOCHAM

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 68 – 70, 2006.

– a compositional rule-based language for modeling biochemical systems, allowing patterns and kinetic expressions when numerical data are available [7];
– a numerical simulator and a non-deterministic boolean simulator;
– a powerful query language based on temporal logic for expressing biological queries such as reachability, checkpoints, oscillations or stability [8];
– a machine learning system to infer interaction rules and parameter values from observed temporal properties [9].

An important characteristic of a language for modeling complex systems is that one may have to consider several semantics corresponding to different attraction levels. It is indeed important to provide the ability to skip from one level of abstraction to another one, and thus to combine several semantics in the language. Perhaps the most realistic semantics is to consider a population of molecules, and consider stochastic simulation as introduced very nicely in the 70s by Gillespie [10]. In Biocham we currently combine two abstraction levels: the molecular concentration semantics and a boolean semantics for reasoning simply about the presence or absence of molecules.

At the concentrations semantics level, a set of reaction rules given with kinetic expressions compiles into a set of (non-linear) ordinary differential equations. At the Boolean semantics, the rules compile into a highly non-deterministic concurrent transition systems which gives an account for all possible competing interactions. The most original feature of BIOCHAM is its use in both cases of a powerful language based on temporal logic to formalize the biological properties of the model. In the (non-deterministic) Boolean semantics, we use the Computation Tree Logic CTL [11], while in the (deterministic) concentration semantics, we use a Linear Time Logic LTL with constraints. We have shown that these temporal logics are sufficiently expressive to formalize a very rich set of biological properties such a reachability, checkpoints, stability or oscillations, either qualitatively or quantitatively[8, 7], and in large models of the cell cycle of up to 500 variables [12]. The machine learning system of Biocham builds on these formal languages and semantics to discover new reaction rules, or fit parameter values, starting from a set of observed properties of the system formalized in temporal logic [9].

Our current work aims at developing a modular modeling discipline for quantitative models based on a full decomposition of interaction rules. Following this approach, we are currently developing a mixed model of the cell cycle and the circadian cycle with applications to cancer therapies.

## Acknowledgments

The work described in this paper is carried out in our group by Nathalie Chabrier-Rivier, Sylvain Soliman and Laurence Calzone mainly. It has been supported by the ARC CPBIO2[2] and is now partly supported by the European STREP project APRIL II3[3] and the European Network of Excellence REWERSE4[4].

[2] http://contraintes.inria.fr/cpbio
[3] http://www.aprill.org
[4] http://rewerse.net

## References


1. Segel, L.A.: Modeling dynamic phenomena in molecular and cellular biology. Cambridge University Press (1984)
2. Chen, K.C., Calzone, L., Csik´asz-Nagy, A., Cross, F.R., Gy¨orffy, B., Val, J., Nov`ak, B., Tyson, J.J.: Integrative analysis of cell cycle control in budding yeast. Molecular Biology of the Cell 15 (2005) 3841–3862
3. Kanehisa, M., Goto, S.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28 (2000) 27–30
4. Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., Karp, P.D.: EcoCyc: a comprehensive database resource for escherichia coli. Nucleic Acids Research 33 (2005) D334–D337
5. Hucka, M., et al.: The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. Bioinformatics 19 (2003) 524–531
6. Fages, F., Soliman, S., Chabrier-Rivier, N.: Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM. Journal of Biological Physics and Chemistry 4 (2004) 64–73
7. Chabrier-Rivier, N., Fages, F., Soliman, S.: The biochemical abstract machine BIOCHAM. In Danos, V., Sch¨achter, V., eds.: CMSB'04: Proceedings of the second Workshop on Computational Methods in Systems Biology. Volume 3082 of Lecture Notes in BioInformatics., Springer-Verlag (2004) 172–191
8. Chabrier, N., Fages, F.: Symbolic model cheking of biochemical networks. In Priami, C., ed.: CMSB'03: Proceedings of the first Workshop on Computational Methods in Systems Biology. Volume 2602 of Lecture Notes in Computer Science., Rovereto, Italy, Springer-Verlag (2003) 149–162
9. Calzone, L., Chabrier-Rivier, N., Fages, F., Gentils, L., Soliman, S.: Machine learning biomolecular interactions from temporal logic properties. In Plotkin, G., ed.: CMSB'05: Proceedings of the third Workshop on Computational Methods in Systems Biology. (2005)
10. Gillespie, D.T.: General method for numerically simulating stochastic time evolution of coupled chemical-reactions. Journal of Computational Physics 22 (1976) 403–434
11. Clarke, E.M., Grumberg, O., Peled, D.A.: Model Checking. MIT Press (1999)
12. Chabrier-Rivier, N., Chiaverini, M., Danos, V., Fages, F., Sch¨achter, V.: Modeling and querying biochemical interaction networks. Theoretical Computer Science 325 (2004) 25–44

# SYMBIONIC: A European Initiative on the Systems Biology of the Neuronal Cell

Ivan Arisi[1], Paola Roncaglia[2], Vittorio Rosato[3], and Antonino Cattaneo[2]

[1] Lay Line Genomics S.p.A., via di Castel romano 100, Roma 00128, Italy
i.arisi@laylinegenomics.com
[2] Scuola Internazionale Superiore di Studi Avanzati, via Beirut 2-4, Trieste 34014, Italy
roncagli@sissa.it, cattaneo@sissa.it
[3] Ente per le Nuove Tecnologie, l'Energia e l'Ambiente
Via Anguillarese 301, Roma 00060, Italy

To understand how genes, proteins and metabolites make up the whole organism, a systemic view is demanded, that is to conceive genes and proteins more as part of a network than as isolated entities. Molecular function becomes then a function of cellular context and not only an individual property. This change of attitude is accompanied by the recognition that bioinformatics plays an indispensable role in extracting information from the huge amounts of data stemming from recent "-omics" research. Such systemic view of cells demands the capacity to quantitatively predict, rather than simply qualitatively describe, cell behavior. In fact, in parallel with the data-driven research approach that focuses on speedy handling and analyzing of the currently available large-scale data, a new approach called "model-driven research" is gradually gaining power. Model-driven research aims at setting up a biological model by combining the knowledge of the system with related data and simulates the behavior of the system in order to understand its biological mechanisms.

The neuronal cell represents a very fascinating and highly complex system. In addition to the basic biochemical processes that are common to all types of eukaryotic cells, such as gene transcription, protein synthesis, and metabolism, neurons are electrically excitable and able to receive and propagate excitation via thousands of synaptic contacts. Traditionally, computational neurobiology has devoted its efforts to model the electrical properties of neural membranes, rather than the intracellular aspects, since the pioneering work by Hodgkin and Huxley in the early 1950's. On the contrary, elucidating intracellular signaling pathways is one of the most fundamental issues in current biology, with important implications for human health. In the last years scientists have realized that the neuronal function could only be truly understood when the morphology of cellular compartments, the location of proteins, and the kinetics of intracellular cascades are taken into account. The post-synaptic component is now viewed as a complex, dynamic assembly of proteins, located in the plasma membrane but also in the cytoplasm. The shape of the neuron is no longer considered to be fixed, but on the contrary processes like dendritic spine remodeling have been recognized as crucial for plasticity events. The overall features of a model neuron could be viewed as the result of the interaction of, at least, six main classes of cellular processes: protein-protein signaling networks, metabolic and enzymatic

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 71 – 75, 2006.

networks, regulative networks, membrane electrical excitation, synaptic communication, and development. All those neurobiological features have to be taken into account if we want to model a whole neuron or even a small part of a neuron realistically. In other words, we have to develop a Systems Biology (SB) of the neuron. Investigating neuronal biochemical processes requires a cross-disciplinary approach, involving on one hand quantitative experimental methods to study excitatory processes, large scale molecular networks and the kinetics of protein-protein interactions, and on the other hand computational modeling of intra-cellular processes and, as far as the synaptic transmission is concerned, inter-cellular communication. Thus, an integration of different experimental and modeling approaches is crucial for a comprehensive description of the cell and for a complete biological understanding of the neuronal behavior. Several scientific and technological expertise are required in order to cope with the great heterogeneity of intracellular processes that must be investigated and described by computational models in such a comprehensive view. We believe that the SB approach is the right one to take into account both the complexity of the neuron and the inter-disciplinary nature of the scientific research in this field.

A comprehensive and quantitative understanding of neurons, in the form of an in-silico model of a prototype cell, shall substantially contribute to the rational design of treatments for human neurological and neurodegenerative diseases. Lay Line Genomics (LLG) is an Italian biotech company, based in Roma and Trieste, whose activity is focused on neurodegenerative diseases. In 2002 LLG promoted an Expression of Interest (EoI) submitted to the European Commission, aimed at launching a large-scale European initiative for the simulation of a neuronal cell. The EoI enabled to coordinate a broad scientific network of research institutions and industries. By then, large-scale projects in the field of SB had already been launched in Japan and the USA, aimed at the computational simulation of whole cells, but not of neurons and not in Europe. Therefore such an initiative would capitalize on the enormous scientific potential of European expertise in cell and molecular neurobiology and neurophysiology, functional genomics, proteomics, bioinformatics, biophysics and computational biology, and would fill a significant gap in the international scientific arena.

Following the initial idea of the EoI, LLG is now project coordinator of a Specific Support Action (SSA) called SYMBIONIC funded by the European Commission within the FP6 (www.symbionicproject.org). Additional funding has come recently from the Area Science Park in Trieste (Italy). The project started in November 2003 and will last 24 months. The three main partners are LLG, the International School for Advanced Studies (SISSA-ISAS) in Trieste and the University of Barcelona. Over 20 other research institutions and industries (in the computing and pharmaceutical areas) from Europe and Israel collaborate to the project. SYMBIONIC is the first step towards the long-term objective of the initiative that was put forward with the 2002 EoI, being the driving force for the creation of a European in-silico exhaustive model of the neuronal cell. This first phase is carried on through a training and dissemination program and several collaboration activities:

- Collaborate and coordinate with other European SB initiatives.
- Train a new generation of young scientists in neuronal SB, both in the computational and experimental fields.
- Disseminate knowledge about the SB of neuronal cell, even to non-specialized audience.
- Contribute to standards.
- Give birth to more ambitious European research and technological projects.
- Raising the awareness of biotech/pharmaceutical and computer industries about the great potential of neuronal SB.

The full title of the project is "Coordinating a neuronal cell simulation initiative with ongoing EU-wide Systems Biology programs", since a key point is to coordinate all the existing efforts towards a broad European SB community, which is also the hope of the funding institution. SYMBIONIC closely collaborates with the SSA EUSYSBIO (EUropean SYStems BIOlogy network, www.eusysbio.org) since the beginning, for the main activities. In particular, joint workshops have been and will be co-organized on the occasion of the International Conferences on Systems Biology (ICSB) in 2004 and 2005, and scientific collaborations were and will be at the basis of the courses on computational and experimental SB organized by SYMBIONIC also in 2004 and 2005.

The SYMBIONIC training program is aimed at forming a new generation of young scientists in the highly inter-disciplinary field of neuronal SB. It is based on two main courses on computational and experimental methodologies for the neuronal SB and on other minor collaborations, in particular SYMBIONIC contributed to a practical course on SB funded by the European Science Foundation (ESF), held in Oxford in September 2004 and organized by the Oxford Brookes University, and it will take part in a lecture course in Austria in 2005, funded by the Federation of European Biochemical Societies (FEBS) and organized by EUSYSBIO. The first training activity organized by SYMBIONIC, a practical course on "Computational SB of the neuronal cell", was held in Trieste in December 2004. The course topics included methodologies to model neuronal shape and development, temporal-spatial properties of molecular networks, synapses, electrical excitation, signaling pathways and metabolism, molecular transport, genetic networks, and sensory transduction. We provided hands-on computer sessions to introduce some of the main softwares and standards used to model biochemical systems. The next course, in fall 2005, will be devoted to advanced experimental techniques for the neuronal SB.

An important event in the dissemination strategy was the ICSB2004 Satellite Workshop on "Industrial Perspectives of Systems Biology" in Heidelberg, Germany, where SYMBIONIC and EUSYSBIO invited key representatives from the pharmaceutical, biotech and academic world to discuss about the role of SB in current research strategies.

The network supporting the SYMBIONIC Action is constituted by a cross-disciplinary set of public and private institutions, each of them bringing specific competences. This is a first important result testifying the fact that the only possible way for a true scientific and technological progress is the leverage on public and private scientific expertise, jointly with a strong application-driven force. This is a major strategic issue: a true mixing between public and private research to strengthen

the European capabilities in the high-technology areas. This will produce a significant fall-out on European industry and increase its competitiveness in the world-wide arena. The presence of a large pharmaceutical company in the SYMBIONIC network testifies, on one hand, the interest of the pharmaceutical industry to reduce as much as possible the time for "technology transfer" from the high-end research into novel application-oriented methods and tools.

A further relevant point that the initial promoters of SYMBIONIC attentively considered is the inclusion of small and medium enterprises (SMEs). High-Tech and Bio-Tech SMEs are very important in the present industrial tissue in Europe, constituting an essential resource of the European economy. Their involvement in the proposed Action can significantly contribute to strengthen their competitive advantage and their paradigmatic role in the economic European context. There are several technological areas where it is easy to predict a sizeable fall-out from Systems Biology initiatives originated directly or indirectly from this SSA. These researches will boost the laboratory activities and generate a consistent quantity of data available to the whole research community.

On a separate matter, the design and modeling of such a complex object as a cell will push at the extreme the actual capability of numerical modeling of complex systems. It will thus produce a strong driving force from other scientific and technological compartments that will be able to transfer specific tools and methods into the emerging area of Systems Biology. This will produce a relevant cross-fertilization, which will represent a major outcome of the SYMBIONIC initiative. Systems Biology will require substantial investments in terms of computing resources and computational strategies. In this respect SYMBIONIC will also address the discussion of future computational strategies (GRID, deployment of large-scale computational facilities, special purpose hardware). The relevance that Systems Biology is assuming in the field of high performance computing, methods and tools is also testified by the interest that this SSA initiative has catalyzed from major hardware companies, that support this initiative. In fact, they predict the explosion of this new area of high-end research where the complexity of models will be certainly accompanied by the need of deploying powerful computing infrastructures, new programming languages and standards for model representation.

All these technologies would result in a series of applications in the biomedical field with a significant impact on human health. A key point is to accelerate the drug discovery and development process, in order to reduce the final costs of drugs and animal testing in favor of computational screenings. The design of reliable in-silico screenings will also increase the rate of success of a candidate compound in the clinical phase. The availability of a comprehensive model of a neuronal cell, where its constituents can be accounted for in a quantitative form, would represent a formidable tool for the analysis of cell behavior under physiological and pathological conditions. Achieving this goal would be of paramount importance, not only for perfecting our understanding of the basic mechanisms of cell behavior, but also to provide a tool for detecting specific components crucially involved in disease. This will contribute to increase the comprehension of the cellular physiology of socially relevant neurodegenerative and neurological disorders such as Alzheimer's, Parkinson's and psychoses.

One of the major expectations of SYMBIONIC participants and at the same time one of the main project aims is to promote and give birth to new and more ambitious research projects. We firmly believe in the idea that the network contributing to the SYMBIONIC initiative is a critical mass of research groups and industries able to launch competitive projects, also in collaboration with other European key players in the field. This is already happening, as some of the SYMBIONIC partners are involved preparing new European projects in the “life science” and “information society technology” thematic areas of the FP6: at the border between these areas lies Systems Biology. Our hope is that the convergence of these actions will lead to a comprehensive and reliable computational representation of the neuronal cell.

# A Biological Approach to Autonomic Communication Systems

Iacopo Carreras, Imrich Chlamtac, Francesco De Pellegrini, Csaba Kiraly, Daniele Miorandi, and Hagen Woesner

CREATE-NET, Via Solteri, 38,
Trento 38100, Italy
{name.surname}@create-net.org
http://www.create-net.it

**Abstract.** Among the most important research topics in computer sciences, a primary role is played by design and control of next-generation communication networks (NGCNs). Such networks will be characterized by heterogeneity at all levels, encompassing a large variety of users, media, processes and channels. Another important feature of NGCNs will be the ability to interact with the environment. Various agents will collect information from the surroundings, and, then take appropriate actions in response, either in a centralized or in a distributed fashion. These features will characterize a pervasive computing and communication environment, a challenging scenario for scientists in all computer sciences-related research fields. Users will be highly mobile, and will need to access services without relying on a end-to-end connection. These factors will reflect into an increasing network management complexity, that will be approaching the limits of human capability.

Consequently, necessary features of NGCNs will be the ability to self-manage, self-adapt and self-organize. These features may be summarized into one single paradigm: autonomic communication (AC). AC is an example where biological systems are considered as models of self-management and self-organization.

This suggests that an appealing approach for governing the complexity of NGCNs is to draw inspiration from biology, as in autonomic computing, in order to achieve an efficient and robust communication system. This requires a multi-disciplinary approach to ICT-related research, which in our view can lead to innovative and creative solutions to the challenges related to next generation networks.

## 1 Introduction

Design and control of the next generation communication networks (NGCNs) are primary research topics in computer sciences. Such networks will be characterized by heterogeneity at all levels, encompassing a large variety of users, media, processes and channels; moreover, different types of agents receive information and transform it into actions in an isolated or cooperative way. This is particularly true as the trend toward ubiquitous computing continues to gain momentum, since in the future

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 76 – 82, 2006.

pervasive environment we can expect the number of nodes to grow by multiple orders of magnitude as tags, sensors, PDAs etc., get fully integrated into the communication superstructure.

Not only will the amount of information in these all-embracing pervasive environments be enormous and to a large degree localized, but also the ambience within which these nodes will act will be intelligent, mobile, self-cognitive, knowledge-based and, in some sense, "almost alive".

The result is a communication network always interacting with the environment, where the heterogeneity present at all levels poses challenges in terms of self-organization and automation of all management processes, since it will not be possible to handle "manually" and all the network complexity all trouble-shooting tasks.

IBM, with the *Autonomic Computing* initiative, was the first to explicitly address these issues [1], [2], [3] explaining that the major obstacle to the progress of information communication technology (ICT) fields is the difficulty in managing today's computing systems.

The properties of self-organization, evolution, robustness and resilience are already present in biological systems. IBM's autonomic ideas, for instance, draw inspiration from the human nervous system. This indicates that similar approaches may be taken to manage different complex networks, which allows the expertise from biological systems to be used to define solutions for governing future communication networks.

Hence, a multi-disciplinary approach to research may represent a viable way to achieve major breakthroughs in a wide variety of problems in ICT.

This article presents relevant examples on how biology has been successfully applied to traditional communication network problems and presents BIONETS, a novel paradigm for the next-generation communication networks.

## 2 Autonomic Systems

The concept of Autonomic Systems is emerging as a significant new strategic approach to the design of computer-based systems and, generally, refers to computing systems that can manage themselves. The importance of this research direction has been recently fostered by the increasing complexity in the design of next-generation self-organizing, context-aware pervasive computing environments.

The general idea is to create systems that are able to behave autonomously according to high-level description of their objectives. In [2], [3], the following basic aspects have been identified as characterizing the self-management of autonomic systems:

- self-configuring: autonomous adjustment and configuration of components;
- self-optimizing: continuous search to improve the performance;
- self protecting: autonomous detection of failures;
- self-healing: autonomous defense against malicious attacks.

Autonomic systems will be generally constituted by *autonomic elements*, a myriad of individual entities that dynamically interact with the environment and other elements. Relations among different autonomic elements need to be specified through

standard and widely adopted languages and ontologies. Really challenging task will be the understanding of the complex relationships among the local behavior and the global behavior, where the system wide issues arise from unpredictable relations and from unpredictable scenarios.

A goal-oriented paradigm is adopted, where humans will be in charge only of the high-level specification of autonomic system's goals. The rest of the work will be carried out by the autonomous systems transparently and without human intervention.

As a practical example, often autonomic elements are seen as agents while autonomic systems as multiagent systems, built on top of a Web service-like infrastructure.

## 3 Related Work

Following the incredibly high similarity between future communication networks and biological systems, an increasing number of researchers are trying to draw inspiration from the living world and to apply biological models to open research issues. The results are *bio-inspired systems*, which are man-made systems whose architectures and emergent behaviors resemble the structure and behavior of biological ones. This research trend, combined with recent technological advances, is now allowing the physical implementation of incipient bio-inspired systems.

Several examples are available in the literature, where biological concepts are considered as models to imitate. Each example focuses on a different biological aspect and applies it to solve or to optimize a specific technological problem. Some of the examples try to mimic the *self-governance* of social and economic systems as well as purely biological ones, while others imitate the cellular reactions and processes.

In [4], [5], principles of the cooperative behaviors of large-scale biological systems have been considered as the example to follow for defining an autonomous evolutionary architecture. The architecture, called Bio-Networking, supports the design and implementation of distributed network applications according to several biological concepts and mechanisms.

The final goal of the Bio-Networking architecture is to exploit the decentralized organization of autonomous biological individuals in order to drive an *emergent behavior*, i.e. the complex global behavior of the network as the outcome of local (simple) interactions among system entities.

In [6], autonomic techniques are applied to system management. In pervasive scenarios indeed, a centralized management approach is unlikely to be adopted, due to the complexity in controlling an enormous number of devices. Drawing inspiration from nature, problems with real-world relevance are approached. For example the autonomic creation of cells macro-structure in animals cells is applied to design the channel allocation algorithm in a cellular network, while a bacterium-inspired software has been developed for the adaptive management of active service networks.

In [7], a new family of FPGAs is first presented, where a fault tolerance mechanism is inherited from the way cells differentiate in a redundant way the instructions coded in the DNA. Then, the human immune system capability of recognizing virtually any foreign cell or molecule is considered as an example to

imitate. A possible mapping is carried out among the immune system and hardware and, finally, a general architecture is proposed in order to reproduce the most important healing properties of cells.

In [8], a self-adaptive distributed agent-based defense immune system is developed and applied to the computer-security domain for detecting and eliminating malicious code or bad packets.

## 4 BIONETS: A Genetic Approach to Today's Autonomic Pervasive Environments

The amount of information in the new emerging all-embracing pervasive environments will be enormous. Current Internet protocols, conceived almost forty years ago, were never planned for these specific scenarios. The communication requirements posed by these protocols on the low-cost sensors and tag nodes are in direct contradiction to the fundamental goals of making these nodes small, inexpensive and maintenance-free. This contradiction may be splitted into two concrete problems: the cost problems and the management problem.

The cost problem refers to the traditional approach of treating all devices as nodes in a network. The vast majority of the devices will be tiny small nodes that will not only be required to sense but also to run a complete protocol stack, including routing and transport protocols. This has a direct impact on the complexity of the nodes, as well as their energy consumption. The use of batteries in tiny devices increases weight and size drastically and places a fundamental limit on their lifetime. The energy gathered from the environment through solar cells etc. will not be sufficient to do packet relaying. As a result, the huge amount of small tiny nodes can not come at a price low enough to justify their mass usage and deployment.

The network management problem refers to the fact that all routing protocols proposed so far for mobile ad hoc networks (which can be seen as predecessor of the future networks) have two major enemies: a huge number of nodes and mobility. Unfortunately, these are the main properties of the networks we envision. None of today's routing protocols will be able to converge fast enough to avoid disconnected network islands. Another problem is the resource management: how to manage the resources of a network (e.g., radio spectrum, energy, computing power) that consists of billions of nodes using hundreds of different technologies? These nodes will have no motivations to waste their energy or other resources in order to relay data packets for other unknown and far nodes.

This situation needs therefore a radically different approach to communication, especially since pervasive and ubiquitous networks are expected to be the key drivers of the all encompassing Internet of the coming decades.

In [9], [10], drawing inspiration from genetics and the rules of evolution, a *service centric* communication paradigm is proposed, which is autonomous, and autonomously self-adaptive. This paradigm is called BIO-inspired NExt generaTion networkS (BIONETS).

### 4.1 The BIONETS Paradigm

Trying to find a better solution to the exposed problems, the BIONETS paradigm totally changes the perspective.

In the new ubiquitous context, the information that is managed and exchanged by users is drastically changing its significance. Information will be constantly localized in space and time, which means that, most of the time, information will simply be outdated and therefore useless with respect to the context where the user is moving in. It will be always possible to define a local sphere (both in time and space) within which the data represents useful information.

Since service provisioning is the original goal of the network, we let the service itself autonomously define how the network is supposed to be in order to satisfy its requirements. Networking will occur only as a consequence of service needs, and the network itself will autonomously evolve and adapt together with the service.

In the envisioned ubiquitous scenario, the environment will dictate the rules of adaptation. Users will be mobile and will change their location over short periods of time, leading to a continuously changing topology. The success of the service will be in its ability to track following these changes and in, subsequently, adapting its main functionalities.

The problem of the management is now shifted at the service level, since now the problem becomes how to provide the myriads of services that users require in a hostile environment, where:

- myriads of services should live and share limited resources;
- the current environment and user needs are so different and changing so dynamically that the service should be able to fine-tune itself.

The key to handle the complexity of the problem is to define a system which is both self-organizing and self-optimizing at the same time: an autonomic system.

For this purpose, BIONETS build the solution on the best and most complex system in this class. Adaptation by evolution is the way how organisms evolved in nature and we propose to apply the rules of genetics to define the service and the process of adaptation. In this sense the network may be interpreted as the *habitat* where organisms are moving and the *genetic information* codes their behavior and goals.

According to the BIONETS paradigm, services are associated with living organisms. The role of the service will be for instance to provide answers to questions like ``How is the weather around the train station?'' or ``Where will I find a free parking space around there?''. Services will be user-situated, meaning that they will be hosted on users' devices and will go around through the physical movement of the users.

Each service is constituted by a program and its related data that are organized into chromosomes. Through the exchange of information among services a *mating* process is defined and offsprings are generated on user's devices. Through mating preferences, some services will have a higher chance to reproduce and therefore to spread. Through mating and mating preferences, service evolves and constantly and autonomously adapts to the environment.

Chromosomes are collections of *genes* that are the smallest service (related) data unit and inborn intelligence/instincts and thus represent all the information needed by the organism to function and by the service to be executed.

As in nature, it is possible to define a complete life-cycle of the organisms and therefore of services. The life-cycle starts from the birth of an organism, goes through the reproduction and ends with the death. Reproduction and evolution occur applying evolution rules inherited from nature.

Fitness measures the correspondence between the organism genetic information and the environment, and determines the mating preferences of services. Therefore, no end-to-end communication concept exists in these systems, and information is only exchanged as needed, locally, between mating organisms. Environment is determining the *Natural Selection* based on the *fitness* of the organisms with the environment and on the mating preferences leading to the best possible services as a function of the environment.

Through this approach, it is possible to define a complete autonomous paradigm for managing services in a completely decentralized manner, where every interaction as well as every decision occurs locally and independently from the rest of the network.

## 5 Conclusions

Modern communication networks are extremely difficult to manage because of the heterogeneity of users, devices, technologies and information. This is particularly true in the new emerging pervasive environments, where almost every object will be integrated into the network.

Self-management is a key requirement for the control of the next-generation networks and autonomic computing the right paradigm to look for.

In defining the basic principles of autonomic communications, nature is the best example to draw inspiration from. In nature it is possible to observe complex biological system capable of autonomous evolution and emergent behaviors.

It is therefore extremely important to foster multi-disciplinary approaches to research, such as the bio-inspired examples we provided in this paper, in order to achieve major breakthroughs in a wide variety of problems. This will probably be the key in transforming the Internet as we know it into the next-generation network.

## References

1. IBM: Autonomic computing: Ibm's perspective on the state information technologies (2001) www.ibm.com/industries/government/doc/content/resource/thought/278606109.html.
2. IBM: IBM and autonomic computing: an architectural blueprint for autonomic computing (2003) http:// www.ibm.com/autonomic/pdfs/ACwpFinal.pdf.
3. Kephart, J.O., Chess, D.M.: The vision of autonomic computing (2003)
4. Nakano, T., Suda, T.: Adaptive and evolvable network services. In: Proc. Of GECCO, Seattle (2004)

5. Suzuki, J., Suda, T.: A middleware platform for a biologically-inspired network architecture supporting autonomous and adaptive applications. In IEEE J. Sel. Ar. Comm. 23 (2005)
6. Shackleton, M., Saffre, F., Tateson, R., Bonsma, E., Roadknight, C.: Autonomic computing for pervasive ICT – a whole system perspective. BT Tech. Journal 22 (2004) 191–199
7. Daryl Bradley, Cesar Ortega-Sanchez, A.T.: Embryonics + immunotronics: A bioinspired approach to fault tolerance. Adaptive Behavior (1996) 169–2
8. P. K. Harmer, P. D. Williams, G.H.G., Lamont, G.B.: An artificial immune system architecture for computer security applications. IEEE Trans. on Evol. Comp. (6) 252–280
9. Chlamtac, I., Carreras, I.,Woesner, H.: From internets to bionets: Biological kinetic service oriented networks. In Szymanski, B., Yener, B., eds.: Advances in Pervasive Computing and Networking. Springer Science (2005) 75–95
10. Carreras, I., Chlamtac, I., Woesner, H., Kiraly, C.: BIONETS: BIO-inspired Next generaTion networkS. In: Proc. of WAC. Volume 3457 of Lecture Notes in Computer Science., Springer (2004)

# The Twilight of the Despotic Digital Civilization

Michel Riguidel[1]

École Nationale Supérieure des Télécommunications, (ENST),
46, rue Barrault, Paris Cedex 13, France
riguidel@enst.fr

**Abstract.** In 2005, we are witnessing the zenith of digital technology. Computers are not going to disappear, but they will no longer be the main motor of innovation. Computers will without doubt make way for nanotechnologies and bio-computing as the motors of progress for western civilization. In the near future, we are going to see an upheaval and further convergence on the horizon: (a) digital convergence before 2010 - resolving differences between industries, (b) nanotechnology and quantum communications after 2010 - an upheaval of seismic proportions and (c) bio-nano cyberspace in the quantum Age around 2020 – a new Era. Despite all its efforts at modernization Europe risks becoming marginalized, incapable, through lack of cohesion, of even partly counter-balancing America's overwhelming domination. This article gives few ideas for a post Internet & mobile 3G society and few thoughts about digital security research & education.

## 1 The Death of Digital Era über Alles

### 1.1 The End of the Supremacy of the All Powerful Digital Technologies

In the middle of the first decade of this century we are witnessing the zenith of digital technology: a precise, deterministic and reliable world, but one that is also vulnerable because it can be cloned, recopied and falsified. It is, in fact, the end of the "all computing" age of Shannon and Turing, the twilight of the information society, which will have lasted sixty years, or the time taken to master the manufacture and multiplication of silicon based transistors. Computers are not going to disappear, but they will no longer be the main motor of innovation. Computers will no doubt make way for nanotechnologies and bio-computing as the motors of progress for western civilization.

The scientific objective of the twenty first century will be to explore the atoms close to silicon, carbon and hydrogen in the periodic table, i.e., the atoms in the middle and upper part of the table. Silicon (Si), carbon (C), hydrogen (H), and oxygen (O) were all mastered in the twentieth century; it is now the turn of indium, germanium, and arsenic, which are light and flexible atoms with even greater degrees of freedom.

In the future, we will no doubt regard the twentieth century as having been painfully slow in terms of research, and as having wasted too much time obstinately producing the heavy elements of the periodic table by artificial means, in an

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 83 – 98, 2006.

unnecessary fixation with radioactivity. A lot of money and energy has been spent exploring the heavy elements. We should forget about uranium and the artificial elements. Einstein's followers are partly to blame for the time we have lost.

In the years to come, we can expect the demise of Moore's Law when we turn the corner into a quantum world. This difficult moment will be full of promise, bringing with it numerous interesting and competing possibilities. As quantum mechanics, chaos, and randomness in architectures and protocols appears, we should seek options that include doubt (true *and* false, true *or* false), probability and stochastic solutions instead of just relying on a distinct binary environment of ones and zeroes. We will no longer have IP packets that travel "optimally" but rigidly through networks. A piece of information that has a probability value can cross a porous or permeable medium stochastically; we can forget the Internet's fixed and simplistic routing, we will be able to send, bits and quantum-bits, packets and quantum-packets, sessions and quantum-sessions across a communication environment (cable or wireless), using porous medium theory rather than over possibly congested routes with a bandwidth.

### 1.2 The Failure of Von Neumann: The End of Omnipotent Software

Von Neumann invented the concept of software. In 1948 he became exasperated with the regular but short-lived repairs (in the form of hardware patches) that had to be made on the architectures of his calculating machines. Instead of applying his intelligence to hardware connections, he preferred to apply it to the contents of the circuits rather than to their topology. Instead of fixing connection errors in his computer with *hardware* patches, he preferred to inject human intelligence and make *software* patches, transferring complexity to the software. However, with time it has become apparent that software patches are now effectively a cancer in computing. It is important to take a new look on this outdated dichotomy between hardware and software. Currently, we put patches on our PC's operating system and our antivirus software each week, so as to update our fragile software configuration. We live in a just-in-time security and dependability society. This way of working is too dangerous. We have yet to master the software manufacturing. There are numerous bugs, and these are due to the procedures involved in programming languages. A computer program is an algorithm that takes up space and time, and because of this there are inevitable side effects.

No doubt, the solution lies in a return back to other more complex hardware. In the future it will certainly be simpler to manage or manipulate the hardware (nanotechnologies) than to make more complicated software. The complexity of reality will be more intelligently apportioned between living cells produced with biotechnology: these are "nanotechnologies", which will have functions anchored in reality. Instead of carrying out simplistic models and simulations, there will be real "experiments" with suitable computer peripherals. The new, possibly quantum, computers will be linked to our current computers for data processing. Computing hardware and software will be classified according to their "hold on reality" and will content themselves with simply managing these peripherals. Mathematical models, too simplistic in their relation to the complexity of reality, will be less valuable.

### 1.3 The Digital Age and Convergence

In parallel to the development of the Internet is the development of mobile telephony, and the use of digital technology in numerous other sectors, to the extent that we now talk of the digital age to describe the appearance of this intangible world that is expanding all around us, this aspect is taking increased value and importance. The value of some businesses essentially lies in their non-tangible assets, made up of data, documents, and software, rather than their premises and equipment, or even their staff.

The convergence of three industries (computing, telecom, multimedia) is far from being complete today. Computing remains in the lead for the exchange of texts (access to text based information on the Web, exchange of texts by email), the telephone remains the main means of communication by voice (the text of SMS messages tends to be in playful phonetic language), and television is still our preferred way of looking at moving images. For the time being only the world of popular music and of the radio stations of adolescents is in the process of being absorbed by the Internet.

### 1.4 The Last Generation of Visible Computers

We should see, in years to come (2005-2010), the appearance of the last generation of nano-processors that will completely invade our everyday environment: disposable computers fixed to objects, pinned to clothes, or inserted under the skin. These are RFIDs, or digital prosthetics. The software for these disposable computers will be accessible via a network, and managed by “aspirators”: readers or scanners that have yet to be invented.

Europe should back this horse, because it has an advantage for us: there are no compatibility issues with the past, everything can be invented: the nano-operating systems, nano-protocols, etc. There is an opportunity to create a new type of computing. It is similar to the start of computing, but without Moore’s Law to spoil everything.

### 1.5 Security of Software and Smart Dust

These latest incarnations of nano-computers will be the last visible computers. After that we can expect the arrival of robots and actuators that will be almost invisible. They will function in groups. It is therefore essential today to begin to use statistical reasoning with computers: a thousand computers carrying out one operation: what happens in this situation? How do we make use of the results? How do we manage, trace and repair them?

There are security issues with software that have never been properly dealt with. Software security is more than just the question of "does it do what it is supposed to do?" (a correction issue). Above all it is about vulnerability issues: “is the software vulnerable to itself, to other software, or a threat to the environment?”.

In the future, security issues will be far more concerned with hardware. With the downsizing or nano-ization, and the “massification” of computing hardware, the question will arise of how to manage the life cycle of disposable and/or potentially threatening computers. Traceability will be more important than ever.

To deal with these future vulnerabilities, digital security must be holistic, taking all factors into account. A technological approach is much too narrow.

The quantum age is unavoidable. Security in a quantum world is an important direction for research: how to build trust and assign security policies using Heisenberg's Uncertainty Principle; how to understand statistical trust, stochastic protocols, and telecom infrastructures based on random geometry. These technologies should give governments a head start, in particular because quantum networks will be a new infrastructure different from the digital ones already in existence.

### 1.6 Shannon and Turing Obsolete: Mobility and Massification

With Shannon it was necessary to optimize the storage and transmission of information, considered statistically with simple, optimal, Gaussian laws, but these hypotheses are worn out; they are no longer valid.

With Turing, computers have been the same for sixty years: a confined machine working with loops and branches ("if then else" and "for i = 1 until n, do this").

These models have now been considered from every possible angle.

Alan Turing had not considered distributed, omnipresent computing. He imagined a single computer, functioning in isolation. Today we have to think in terms of "ambient" computing, with millions of computers cooperating to perform an operation (such as calculating prime numbers, policing the Internet, or regulating the earth's pollution). We must therefore think of computers functioning together. This poses significant theoretical problems that are considerable.

Current grid computing projects are far from dealing with this issue and have to make do with middleware solutions to achieve the asynchronous cooperation of only a few computers.

### 1.7 An Operating System, Not Hunched Up but Open and Part of an Ambient Intelligence

Today's computers use archaic operating systems: Linux, Unix, which was invented thirty years ago (and is designed for hosting servers), and Windows, which is twenty years old (a computer running DEC's VMS for a user).

All of these operating systems manage space (memory and the hard drive) and time for a short period (scheduling that manages task priorities a few milliseconds ahead of time). From time to time these systems deal with input and output.

Such systems keep the external environment at arm's length, and are effectively functioning in a withdrawn state. This was normal thirty years ago, because the emphasis was then regularly on the speed of computing; today this is no longer important, since computers do more than just computing.

We therefore have to return to the basic issue and spend computing energy where it is most needed. No matter that timing or memory is not optimized, what is important is the environment, or ambient intelligence. What we need is a computer that has been designed with access to the outside world as being the priority. The management of input and output (the multiple inputs of networks such as WiFi, Bluetooth, GPRS, etc.), the mobility of the computer (interconnection, interoperability), its flexibility (the possibility of reconfiguring itself to save energy, or to take into account changes in the environment), and security (is an environment threatening) must be integrated

into the heart of the operating system. The possibilities of grid computing, which needs an OS which virtualizes material resources, are part of this issue.

### 1.8 Digital Durability

We have just barely digitized museums, administrative data, etc., and it is time to change the physical format. The formats for databases have no continuity: we need to think of ways of storing information permanently. We need to be sure that archives will be easily accessible in thirty years time. The management of the digital life cycle is an unresolved issue. Currently the volume of information stored is doubling every year, i.e., even if information is duplicated hundreds of times and most of it has "fossilized" we still store as much information each year as all previous years combined.

The management of personal digital assets is also unresolved, at individual, business and administrative levels. The massification of digital data brings with it all sorts of "pernicious" problems: congestion, waste, management of archives, etc.

### 1.9 Research Topics

We need to work on facilitating change; the main difficulty of ICTs stems from the fact that they have trouble in assimilating the past. The environment has become suffocating, with any new idea brushed aside by conservatism (in the form of the IETF or the hegemonic IT industry).

We have to work on the management of large computing structures and on quantum programming languages and on apprehension over nanotechnologies.

We have to work on the fact that we are currently witnessing a "sedimentation" of technologies and that a digital system is like a geological outcrop, that presents the designer a succession and aggregation of technologies, standards, and anachronistic paradigms and that we have to change these systems today, taking into account this temporal perspective.

## 2 Information Architecture on the Web

### 2.1 Introduction

To understand and put search engines into context, one must, on the one hand, consider the architecture of the Internet, the structure of the Web, access to information, the philosophy and ideology underlying the Internet, and the architectures in use, and on the other hand, the role of the Internet in the current process of digital urbanization.

There are two key factors in the technical success of the Internet: its protocol and its information access service, the World Wide Web. The messaging systems had no role in this success, and were existent before the advent of Internet.

### 2.2 IP Protocol and Internet Architecture

Communication protocol (IP: Internet Protocol) is crude but efficient and enables information to be sent from one place to another quickly and asynchronously without

following particular routes. Easy interconnections between different sites and between different routers have been the driving force behind the Internet's expansion. Initially the linkage of the Internet was redundant and the network presented itself as a robust net.

As it spread out, the infrastructure was subjected to the laws of economics and was optimized in order to encourage large amounts of traffic. This resulted in the detriment of the initial idea of robustness. The original design had to be replaced with something that could grow with the volume of traffic, much like a national road infrastructure with its highways and local roads. This very flat architecture facilitates systematic exploration of the Internet by robots or "spiders" that return every month to the same place to collect, update and index information.

### 2.3 The Structure of the Web

The Web has an equally simplistic architecture. Web sites are made up of autonomous pages, and the architecture of a site is no more than a collection of pages written in HTML. It is one of the crudest markup languages. Its semantics are rudimentary. Its success stems from its capacity to encapsulate small applications written in a programming language (Java). As Java was created for small programs at around the time of the birth of the Internet it was chosen as the programming language for the Web.

The Web is like a unique book that an evil demon has ripped up and scattered its pages all over the planet. Search engines are thus a way of reconstituting a part of this ever changing virtual book as it exists at a given moment in time.

### 2.4 Information Access as Part of Everyday Life and the Need for Search Engines

The Web and Internet protocol have favored the flattening and uniformity of computing architectures. Information access has become part of everyday life in the form of heterogeneous text pages, destroying in the process any strange computing constructions and leaving any complex literary constructions to one side. This kind of raw information, without any form of authentication, leaves room for mountains of lies and falsehoods from anonymous writers hidden behind virtual sites. With its success the Web has become a way of making personal information available to everyone. The internet is increasing in size because of its protocol, which produces a snowball effect. The number of sites has grown so quickly that lists of sites have become unusable. More sophisticated ways of searching the full text of sites have become necessary in order to know about their content. Their simple structures facilitate systematic exploration by computing tools which suck up site content and then create a text based index. The initial wasteland of data has made these search engines indispensable; they are the only means of sorting and making inventories of everything that is available on line.

### 2.5 Network Ethics and the Cut-and-Paste Philosophy

The Internet also came with its own set of ethics, based on pragmatism: its designers were champions of liberty and everything was meant to be free. This poorly thought

out philosophy quickly produced an environment somewhat reminiscent of the jungle. The scale of the Internet (to be on the Web is to be visible all over the planet) is, most of the time, an illusion. Cultural roots and national preferences are in fact very well established. However, duplication, imitation, and mimicry have become the norms, destroying the possibility of real diversity, and largely hijacking creativity. Those entering the market face massive competition and the survivors generally share about 80% of the market. This is true for search engines, computer processors, routers, operating systems, word processing packages, databases, virus protection etc. Instead of an open world of free information we are witnessing a digital world turning in on itself with a cut-and-paste philosophy.

Some people believe that everything is on the Net, that you have to be on the Net to exist, and that you can create anything by using what is already on the Net. Those people believe that all computing programs have already been written and are available on the Net. Consequently it is unnecessary to know how to program, you just have to stick a few programs together in order to create a new service and make a fortune…They consider that all reports have already been written, all views expressed, and all are available on the Net. So there is no need to come up with anything new for a dissertation, it is faster to surf the Net and assemble the document that you would have written, but which someone else has already written better…

The cut-and-paste mentality is now so widespread that some university professors even have software programs that can measure the degree and complexity of cutting and pasting that has gone into a student's dissertation. An evaluation of a dissertation thus has to take into account not so much the content as the student's level of comprehension and assimilation of the cut and pasted material.

### 2.6 From Client Server Architectures to Intermediation Architectures to Peer to Peer

The way that networks function has changed. In the nineties client server architectures, i.e. direct links between the supplier and receiver of information, were standard.

Towards the end of the nineties and the beginning of the new millennium intermediation architectures appeared, i.e. architectures in which everyone was both a supplier and receiver of information, offering and searching for this information through the network, formatted according to simple norms. To find anything on this network it was necessary to create mediators such as search engines and the numerous other "services" that mushroomed during the internet boom.

Today we are seeing a further mutation with peer to peer architectures in which the trend is to exclude these intermediate servers and go for direct interaction. Examples are the architectures used to exchanges music files. It is certain that with new uses of the Internet and of mobile telephony, other engines and other architectures will be created to help people find each other, or to hide, in the ever expanding labyrinth of sites.

### 2.7 The Massive Bias of the Google Infosphere [1]

Google is a way of getting an instantaneous response to questions of a dictionary or encyclopedic nature, or of skimming through what has recently been written on a certain subject.

In forming an opinion about something, or of reflection on complex issues, search engines provide too much information and we end up wasting time. It is more convenient to look at a book from a traditional library. To learn mathematics or computing, nothing beats a book with exercises. And courses on the Web in American slant English, and may not be palatable to other cultures.

Viewing of information on the Internet is heavily biased: 95% is in US English, 1% in French, virtually nothing in Chinese; 75% of sites visited are North American, and have a bias towards the computer industry, marketing or e-commerce. Health and law are under-represented. Literature, philosophy and psychology have only token representation. On the Web a new laser printer is more important than the entire works of Aristotle, Cervantes, Shakespeare, Goethe and Victor Hugo; the most recent trend in music passing will create more traffic than all the on line searches relating to cancer since the birth of the Internet.

### 2.8 A real Time Indicator of the Preoccupations and Subconsciousness of the Online World

Managing a search engine is an excellent way to learn about the preoccupations and even the subconsciousness of the people connecting. As such, it is an enormous source of information. A search engine can be used as a means of electronic surveillance (e. g. to locate pedophile networks), or to gather economic intelligence (which companies are being searched for and why). It is sufficient to install a selective monitoring system, suitably suppressed so as not to slow down the search engine. When search engines extend to mobile telephony and to personal objects, they will become even more powerful tools for intruding on digital privacy.

### 2.9 Fungible Information

Most of the information that one finds on Google is of a fungible, generic, run-of-the-mill, neutral or, more precisely, politically correct nature.

For example, one can find good computing courses on the Internet. For the time being, however nothing has really replaced traditional courses with flesh and bone teachers who can spontaneously interact with their students.

There are plenty of parallel virtual worlds of information but for now, these are only negligible niches in the on line environment, and in terms of size and traffic they are obliterated by the weight of the market and institutions.

### 2.10 Intellectual Property, Loss of Identity and the Utopia of Universal Identity

One of the attractions of the Internet is that one can hide there and remain anonymous. Thus one can also steal someone else's identity or use a pseudonym and this has resulted in all the viruses and spam that plague today's environment.

This logic of anonymity on networks has been a major factor in the growth of illegal digital copying and unauthorized imitations, activities made extremely easy by a fundamental feature of digital data, which is that the content is independent of its physical format.

Due to the structure of the Web, we are witnessing much loss of identity. Widespread use of the Web encourages the disappearance of authorship and of

intellectual property. Internet's surfers replay the masterpiece of Luigi Pirandello *Six Characters in Search of an Author*.

Identity can only have real meaning in a closed and ultimately quite small world. When a world is infinite one can no longer name things. For some this will no doubt bring to mind Zermelo's Set Theory which answers the question: is it possible to choose one element out of an infinite number of elements?

We are now asking this question again with the extension of IP addresses: some would attribute an address or group of addresses to everyone and to everything. Others ask this question in the context of smart tags (RFIDs) which will replace bar-codes: instead of attaching a code to objects we will now attach a computer chip. This will enable objects to communicate at say supermarket counters, toll roads, etc.

The logical extension of this concept is to have tags for all objects and a universal communication network. This would be a nightmare in terms of security: hackers would be able to manipulate the tags and pass off a carton of yogurt for a car; it would also pose a threat to privacy, because these tags could become public and relinquish confidential information to indiscreet readers of nearby tags. We risk tagging everything. Human tagging is already underway: a digital device is inserted under the skin of some amnesiacs in order to monitor their location.

### 2.11 Exponential Growth of Digital Assets

The quantity of information in the world doubles every year. This growth is almost certainly related to the price of storing information, which is currently halving every year. This means that statistically each one of us is storing as much information each year as in all previous years combined. This pace cannot continue for very long: we will eventually accumulate so much digital fat under our belts that we will be unable to move.

The information on our computer's hard drives and on servers is unprocessed, often obsolete, and difficult to update. There are no mathematical theories on the aging of information that can deal with it when it is becoming obsolete. Software for summarizing information on a disk or network is not yet sufficiently advanced. A lot more research in linguistic engineering will be required if progress is to be made in this domain.

The quality of a search engine often lies in its ability to determine the recentness of a web page and to give information a temporal and geographic reference (such as location dependent search engines); this is ultimately a negation of the Internet's original virtualization.

### 2.12 Vacuous Information

One must consider that a human is unlikely to write more than a few hundred megabytes of text during his or her entire life. The works of Johann Sebastian Bach fit on a small hard drive. The largest encyclopedia takes up a few gigabytes.

Most of the information on the Web is thus redundant, entropic, often heterogeneous and semantically poor. The works of Proust and Céline, of Lacan, Foucault, and Lévi-Strauss are not even available in their full versions. A complete history of the twentieth century is not on the Internet. The information available

through search engines is that which appears on the main sites having high volumes of traffic. It is no more than the froth of the exchanges in this ocean of words. The search engines have very little linguistic intelligence, and even less semantic intelligence; they are statistical ogres which, having sucked up strings of characters, collect lifeless words and the geometry of links. The meaning of words escapes search engines. The essence of these devices is no more than a clever conjuring trick: statistical analysis of the proximity of words creates an environment which provides a certain kind of meaning. More subtle “text mining” is required to find real nuggets of information. The standard search engines are not capable of this.

### 2.13 Digital Privacy Violated Without Our Knowledge

There exists today a real problem in the general level of digital privacy.

Our personal digital assets are the data that we have stored on our personal computers, our PDAs, our mobile phones, and our credit cards. We access and manipulate this data of our own free will. But there is also all the personal data that is managed without our knowledge. This is the data held by employers, medical files maintained by health administrations, and localization data held by telecom operators who know the location of a mobile phone as soon as it is switched on, photos taken by surveillance and speed cameras, voice messages that are recorded on networks, thereby leaving the user's biometric identification.

The center of gravity of our digital assets has moved out of our control. Our personal data is like a plant root system that has spread creating new “bulbs” that generate further systems. There is even information about us on Google. The infosphere extends beyond our reach, and there is no longer any hope of controlling its rhizomorphic (root like) architecture.

## 3 Upheaval and Further Convergence on the Horizon [2]

### 3.1 Digital Convergence Before 2010 - Resolving Differences Between Industries

Networks make up the visible and invisible nets of transmission and communication. There is now an environment in which all personal digital objects, mobile and fixed, are connected to a network that is always accessible. During the excitement over the Internet bubble in 2000, we often heard of the promise of digital convergence, which is in fact a double convergence of the carriers (the fixed and mobile infrastructures) and the content (voice and data). The standardization of networks will ultimately result in the fusion of the previously distinct computing, telecom and audiovisual industries. This configuration will produce a sophisticated combination of next generation Internet access and fourth generation mobile telephony, the success of which will depend on user adaptation. We will thus see an evolution of networks from vertical segmentation by infrastructure to horizontal segmentation by type of service. Eventually we will have “seamless” global interconnection between telecom operators and the Internet. This convergence is hindered by bitter antagonism between the three industries over any alignment. Of course, the computing industry, the largest “tectonic plate” of the three, will win and convergence will take place through a consolidation of architectures and a realignment of computing standards.

Telecom and audiovisual services will thus move into a less reliable environment, similar to the world of computers that we know today with its accompanying viruses, spam, thefts of identities, etc. The immediate work of digital security research is to restore trust in services that have been "migrated" as a result of the technological convergence of standards and architectures in these industries. This phenomenon of gradual mutation may be worrying in terms of security, but heralds a new era of openness and richness of exchanges and applications for individuals, businesses and the state.

### 3.2 Nanotechnology and Quantum Communications After 2010 - An Upheaval of Seismic Proportions

Turing's universal computer was the dominant machine of the second half of the twentieth century, and its basic construction has not changed since its creation. The power of computers has doubled every eighteen months, in accordance with Moore's Law, progressing from four thousand transistors in 1970 to a hundred million in 2000. Miniaturization will be followed by techniques enabling finer and finer processing technology. We will soon see the last generation of traditional computers, which will be finger-sized rather than hand-sized, and will become part of the fabric of daily life. Smart tags will replace bar codes, and disposable computers will be inserted in objects, clothes, and under skin … These miniaturized communicating objects, linked to the Internet, will offer a virtually unlimited selection of services. The software on these disposable computers will be accessible from public terminals and managed by devices that have yet to be invented. At this point, we will have pushed transistors to the limits of their atomic scale, but will have entered the age of nanotechnology and traditional computing will make the leap into the quantum age. Nanotechnology will enable the creation of intelligent nano-devices, capable of carrying out specialized tasks. They could, for example be used in a telephone microphone. Quantum cryptography will ensure the transfer of confidential information contained in a new generation of credit cards and SIM cards.

### 3.3 Bio-nano Cyberspace in the Quantum Age Around 2020 – A New Era

We can already foresee another convergence that should take place around 2020: that of nanotechnology, biology, quantum communication and traditional computing. This future confrontation is likely to be fierce, and it is a fairly safe bet that traditional computing, with its dreary ones and zeroes, will not survive. This will be the end of the information society.

This will also be the undoing of digital computing as we know it (Shannon's coded data and Turing's coded programs).

Hardware, locked in the solitude of Moore's law, will have to break the chains of its obsession with ever finer processing technology. Computing, the prisoner of a double asymptotic curve, consisting of a monotonous dead-end in hardware and inextricable complexity in software, will have to escape the current deadlock if it is to forge the post Internet society.

Computing and Telecom must reconnect with physical reality, returning to the atoms close to silicon in Mendeleyev's periodic table (the carbon and hydrogen of

living biological cells, indium, germanium, and gallium) and interfacing their peripherals with the human sense organs in order to ensure continuity with our biology. Mathematical modeling and computer simulations will ultimately generate hybrid applications. They will be computers in terms of data processing, but they will also be "real" in the way they will be made up of *in situ* experiences.

The telephone of the future is no doubt the first device on this new horizon. Words will be transmitted continuously from our larynx to relay stations and reception may even by-pass the eardrum. This will be a return to the analogue world: computers will draw on various sources, with biological, nano and quantum peripherals. This new world will reduce the Joule effect of circuit heating: computers, telephones and PDAs will have peripherals with motors made up of autonomous atoms, communications and data processing will use quantum states of matter, and bio-computing human-machine interfaces, and the totality will be much less polluting than what we have today.

All the ontologies of this new world, these clusters of living cells linked to a computer, these fleets of nano-robots, and these bundles of photons will be distributed in networks. Some will even be carried within the human body, however, for the time being any representations of this kind environment remain in the realm of science fiction.

## 4 Restructuring Education and Research in Digital Security [3]

### 4.1 Education and Research in Digital Security

Digital security is not taught as a discipline in its own right. In research it is associated with traditional disciplines: mathematics for cryptology, signal processing for audio-visual watermarking, electronics for hardware, computing for the security of operating systems, networks for Internet protocols, telecoms for GSM security, etc. The uses and the sociology of trust are dealt with in Human and Social Sciences. The economic and geostrategic aspects remain virtually untaught. It is essential to create a transdisciplinary domain. The current division into scientific departments goes all the way back to the transformation of European universities by Humboldt in 1809.

No one is unaware of the real stakes in the domain which concerns us; no one is unaware of the real stakes:

- Is France (and Europe in general) to abandon research on the foundations of ICTs and accept a position as a second class player? In France research on modern computing theory has been practically nonexistent for fifteen years and the computing industry has never properly existed, despite numerous initiatives over the past thirty years. Digital security is suffering from this lack of industrial activity.
- Are we to content ourselves with applied research and innovation in the field applications (education, health, transport, leisure, etc.), in services and in uses? Are we to be marginalized and left with a minor role in computing applications? (Digital security comes in at the foundation stage of digital structures and cannot be built on the moving sands of applications.)

- Are we to go along with the general direction of network infrastructures, giving our remarks free of charge to international industry, criticizing propositions through the IETF and international work groups (Liberty Alliance, etc.)?
- Are we to cut ourselves off from innovation? Taking a back seat role in the security of computer applications and services, and multimedia security with foreign computer equipment, much as in computing (computing applications, simulations, and virtual reality).

The sovereignty of France and Europe is in danger: are we to accept this strategy of withdrawal and the loss of any technical autonomy in these sectors?

It is vital to create a critical mass in research, in proportion to the size of large European countries (France, Germany, UK, Italy, etc), for a more clear vision. This critical mass will then serve as a springboard for research in Europe.

The direction of research in digital security must take a long term view and should correspond to a general scheme for the handling of regional, national and international crises.

- Future crises will be marked by mass terrorism attacking the value systems of democratic society (that of the eighteenth century philosophers).
- Future cyber-crime will increase and take more pervasive forms in order to finance the traditional mafia activities (trafficking in narcotics, money laundering, prostitution, secret funding of various groups, etc.).

### 4.2 Recommendations for Restructuring and Energizing Education and Research in Digital Security in France

A field trip to Washington [3] took place in mid March 2004 to study the teaching and research situation of computing and network security in the United States. The following recommendations and reflections resulted from the trip:

- Digital security is a recognized discipline, taught in higher education in the United States.
- Digital security is a research domain that is extremely well supported in the United States by the Department of Defense and the NSF.

At a moment when teaching and research are the subjects of debate in France, it is important to properly identify the stakes and to be aware of the context in which these fledgling interdisciplinary domains must be viewed. The difficult situation of research in France today, the reform of higher education on a European scale, the development of European research programs, the lively international competition, and the potential disruption of international threats, are all reasons for deep reform and a fresh start in the area of security for digital structures. A strong link between teaching and research is essential for the growth and dissemination of knowledge, and for a beneficial interaction between these two activities.

### 4.3 Digital Security in France

At a time when France is about to vote on a law concerning the direction and programming of research, the following recommendations should be considered:

- Digital security is a high priority multidisciplinary activity for the country.
- Under the guidance of a ministry yet to be defined, an initiative should be launched to promote teaching and research in digital security. The steering committee for this teaching and research would include technical colleges and universities, private-public and civil-defense representatives.

It would be advisable to create a completely new high level branch of teaching and research in digital security in France.

1. This branch could initially be formed by an across the board unification of the current programs and courses of the big higher education establishments (ENS, Mines, Ponts, Telecom, ENSTA, University Paris 6, etc.).
2. The teachers of these various schools would be coordinated under a single high level teaching program, adapted to individual applications and to the nature of each establishment, and the program benefiting from the best teachers. This reform could be part of the restructuring of the LMD.
3. Courses would be given to students at technical colleges and to undergraduate and graduate students at universities. These would be “L” and above "L" students in the LMD Bologna agreements. They could thus obtain a certain number of credits (ECTS) within the context of personal study programs.
4. The content of this teaching and research would cover the theory and applications of digital security. Students would have to have already reached a basic level in mathematics, computing and networks. The program would include Cryptology and cryptographic protocols, Steganography and watermarking, Quantum security, Computing and networks security, Technologies for the protection of privacy, Management of digital rights, National digital security, Protection of critical infrastructures, Robustness of systems, Crisis management, Digital government and ethics, Security and privacy of personal digital assets, Digital trust, Methodologies, assessment and certification and Human aspects.

It would be advisable to revitalize fundamental and applied research and innovation in digital security. The collaboration and scope of this research would be on a European and international scale.

On the research level all those involved in digital security research, and who are currently spread amongst the various national research institutions (CNRS, INRIA, GET, etc.), should be united in a Virtual Laboratory.

- The funding of this Virtual Laboratory would come from various ministries (Industry, Research, Defense) with corporate participation (Thales, EADS, Sagem, Alcatel, France Telecom, Gemplus, etc.), possibility via a research Foundation.
- This Super Laboratory of Excellence would receive European funding through participation in European research. It would also collaborate with North American research bodies.
- The integration of engineers into industry will be facilitated by the issuing of diplomas from this laboratory.
- Research subjects would be the same as those listed above for teaching.
- Partnership between businesses and research bodies is vital in digital security.

The internationalization of research is essential for the protection of digital urbanization and transnational infrastructures. Greater visibility and clarity within research in digital security must be the prime objective of this Virtual Laboratory:

- The Laboratory should collaborate closely with the United States.
- A collection of scientific reviews should be created and made accessible on line.
- A consolidation of the congresses, symposiums and conferences should take place in order to improve collaboration, particularly in France.

The above proposed recommendations, if they can be carried out without excessive delays, should improve the visibility and efficiency of higher education and research in security, without challenging the legitimate concerns of teachers and researchers or upsetting budget figures, because the above propositions are aimed essentially at redeploying structures and resources that already exist, even if they are widely dispersed.

The creation of a Virtual Laboratory for digital security may be emerged as a bold yet slow approach. In the short term, some voluntarist measures could be taken, for example reinforcing the subject of security in various national research and innovation networks and encouraging the launch of a common "security" policy for all of these networks. These networks favor partnerships between public academic research and private industrial research and are likely to cover all the issues linked to digital security. The multidisciplinary topical network on security of the *Information and Communication Science and Technology* (STIC) department of the National Center for Scientific Research (CNRS) could be made more operational and dynamic, and be structured more openly. These measures must obviously carried out with the agreement of those involved, in consultation with the STIC Department of the CNRS, and be part of discussions about the future of the networks for innovation and research in technology that have been beset with major funding problems since 2003.

Apart from these national initiatives, the value of collaborating with digital security research laboratories in the United States is beyond any doubt. Concrete action should be taken immediately to initiate collaboration between the United States and all of the relevant bodies in France.

## 5 Conclusions: Research in General, Digital Science in Particular

Research has been a controversial subject in France in 2004, not only because of its content but also because of its organization and funding.

The debate in the public arena has highlighted the impoverishment of research (the modest salaries of researchers, low investment, and massive reductions in the number of research positions), and the cumbersomeness and number of existing bodies (researchers are often hindered by their status, have difficulties at the end of their careers and lack of mobility).

The controversy has tended to ignore the priorities of research (the difference in volume and allocation, in terms of research and effort, between such fields as biology, nanotechnology, information technology and communication), its quality (how research or a researcher is evaluated?), its use (how is fundamental and applied research, and innovation actually applied, what are the links with industry?) and its

content (what should we be researching in France, in Europe and in the world? What are the priories for France and Europe?).

Conservative forces and old-world interests at all levels have hindered genuine discussion. Moreover, the segmentation on which these debates hinge is out-dated (consisting of conformist brotherhoods of physicists and biologists, and corporate institutions). Discussions are biased and reach foregone conclusions (as a result of the lobbying of powerful research bodies keen to retain their position, and rearguard fighting between “Grandes écoles” and Universities), to such an extent that there is little room for calm, objective debate.

If nothing changes, then we may witness something of a decline in the years to come as Europe digests its new member countries. Despite all its efforts at modernization, Europe risks becoming marginalized, incapable, through lack of cohesion, of even partly counterbalancing America’s overwhelming domination. Research collaboration in Europe is not easily developed. The dispersion of research activity over the other European countries does not encourage the creation of synergy.

The political decision-makers and institutional officials for the renaissance of higher education and research must nevertheless redouble their efforts to provoke a drastic overhaul and a restructuring of education and research for the new century. We must not be content with fainthearted gestures, with structures and organizations superficially reinventing themselves, and with measures designed for their media impact rather than for their effectiveness.

It is strange that Digital Science has not yet achieved noble status in the scientific Pantheon. Digital science does not even exist in its own right. Its position is only considered relative to the other “true” sciences (mathematics, physics, astronomy, medicine, the human sciences, etc). We talk of networks and computing, or there are catch-all acronyms like the “ICT” (Information and Communication Science and Technology).

Furthermore, enlightened minds frequently complain and make summary judgments on these subjects, of which they are not well-informed and that they often underestimate. Some intellectuals associate the Internet with technological fetishism, while others play down the importance of networks, saying that, “the Internet has not been behind a single scientific discovery”. This litany continues: there is no shortage of such damning remarks, which are slightly pathetic in their lack of perceptiveness.

## References

1. Riguidel, M.: La sécurité à l’ère numérique, Les Cahiers du numérique, Volume 4, n°3-4, 2004, N°ISBN 2-7462-0907-1, Editions Hermès Lavoisier, Paris (2004)
2. Riguidel, M.: Le téléphone de demain. N°ISBN 2-7465-0209-7, Editions Le Pommier/Cité des sciences et de l’industrie, Paris (2004)
3. Mission pour la Science et la Technologie, Ambassade de France aux Etats-Unis : http://www.france-science.org/photos/1089105199_R-securite.pdf.

# A Compositional Approach to the Stochastic Dynamics of Gene Networks

Ralf Blossey[1], Luca Cardelli[2], and Andrew Phillips[2]

[1] Interdisciplinary Research Institute, Villeneuve d'Ascq, France
[2] Microsoft Research, Cambridge, United Kingdom

**Abstract.** We propose a compositional approach to the dynamics of gene regulatory networks based on the stochastic π-calculus, and develop a representation of gene network elements which can be used to build complex circuits in a transparent and efficient way. To demonstrate the power of the approach we apply it to several artificial networks, such as the repressilator and combinatorial gene circuits first studied in Combinatorial Synthesis of Genetic Networks [1]. For two examples of the latter systems, we point out how the topology of the circuits and the interplay of the stochastic gate interactions influence the circuit behavior. Our approach may be useful for the testing of biological mechanisms proposed to explain the experimentally observed circuit dynamics.

## 1 Introduction

Within the last years a general consensus has emerged that noise and stochasticity are essential building elements of gene regulatory networks. A quantitative understanding of their role is thus needed to understand gene regulation. Regulatory functions can indeed work to eliminate stochastic effects [2], or to even exploit them [3].

In line with new experimental techniques to measure and quantify such behavior, efficient ways to model and simulate gene networks need to be developed, which are currently lacking. Simulations based on differential equations for the concentrations of the various biomolecules, the long-time standard of modeling in biochemical systems, are not well suited for this purpose, except in particular cases. Stochastic effects, which are typically important when molecule numbers are small, are difficult to build into such approaches, and the resulting stochastic equations are time-consuming to simulate. In addition, differential equation models are inherently difficult to change, extend and upgrade, as changes of network topology may require substantial changes in most of the basic equations.

In this paper, we follow a different route. It has recently emerged within computer science in the context of process calculi, and their applications to biological systems. Process calculi [4] are essentially programming languages designed to describe concurrent, interactive systems such as mobile communication networks. Among the various process calculi, π-calculus is one of the best studied because of its compactness, generality, and flexibility. Stochastic variants have appeared recently that address biochemical modeling [5]; they have been used to model molecular interactions [6][7], compartments [8][9], and metabolism [10]. A remaining challenge

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 99 – 122, 2006.

is to model gene networks, to fully demonstrate the flexibility of process calculi, and to eventually support the integration of molecular, gene, and membrane networks in a single framework.

Here, we introduce process calculi by example, in the context of gene networks; technical details of the approach can be found in the Appendix. Modeling with process calculi is very much like programming. It is carried out in concurrent, stochastic programming languages that can easily support very complex and detailed models in a modular ("compositional") way, where separate program units correspond to separate biochemical components.

Our purpose here is in part tutorial: we aim to show that we can do things *simply* to start with, and already get interesting insights. Models in which molecular details are explicitly treated can be built when needed; e.g. see [11] for a discussion of transcription-translation in phage lambda. In addition to our approach being on the level of gene gates rather than molecular components, we have chosen a style of presentation which we believe will be helpful to researchers from neighbouring disciplines (physics, mathematics and theoretical biology), for whom the existing literature on the application of the stochastic π-calculus may be too demanding.

The paper is structured as follows. We first explain how to represent gene network elements as processes in the stochastic π-calculus and how to execute them. We then apply this representation to model gene networks of increasing complexity, and study some of their behavior. In particular, we address the repressilator circuit [12] and two of the (still controversial) examples of combinatorial circuits first discussed in [1].

## 2 Modeling Gene Network Elements

### 2.1 Nullary Gates

We begin by modeling genes that have constitutive transcription but no regulatory control. We focus on the *actions* that are involved in the functioning of genes and molecular components. The generic term *process* is used for any mechanism performing actions and thus progressing through distinct states.

A nullary-input gate (Figure 1), given by a process written *null*(*b*), has a single parameter *b* that represents its transcription product; it takes no input from the

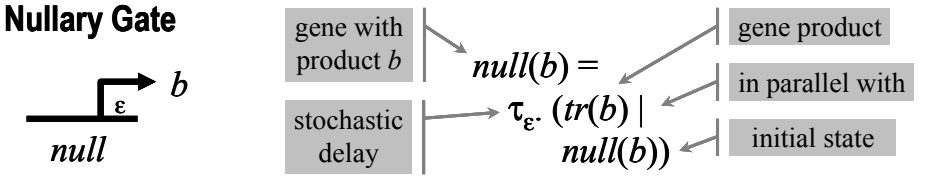



**Fig. 1.** A gene, *null*(*b*) with constitutive transcription, but no regulation (nullary). The product is a translated protein, *tr*(*b*) that attaches to a binding site *b* on some other gene; the definition of *tr*(*b*) is given later. The definition of *null*(*b*) says that this gate waits for a stochastic *delay* ('τ') of rate ε, and *then* ('.') evolves into two processes in *parallel* ('|'); one is *tr*(*b*), and the other again *null*(*b*), the initial state.

environment. The initial action performed by such a gate is a stochastic delay, $\tau_\varepsilon$, where $\tau$ is a symbol indicating delay and $\varepsilon$ is the stochastic reaction constant, which gives the probability per unit time that the delay action will occur [14]. In general, each action in the $\pi$-calculus is associated with a corresponding stochastic reaction rate, such that when an action with rate $r$ is enabled, the probability that it will happen within a period of time $t$ is $F(t) = 1\text{-}e^{-rt}$ [15]. This distribution exhibits the memoryless property, as is required for the Markov property of the stochastic dynamics.

After such a delay action, the original process *null*(*b*) becomes (i.e., changes state to) two separate processes in parallel (separated by the operator "|"): *tr*(*b*) and *null*(*b*). The second process is a copy of the original process *null*(*b*), which was consumed when performing its initial action. The first process, *tr*(*b*), described shortly, represents a molecule of a transcription factor for a binding site *b* on some gene. All together, the *null*(*b*) process is defined as $\tau_\varepsilon$. (*tr*(*b*) | *null*(*b*)). A stochastic simulation of a *null*(*b*) process on its own produces multiple copies of *tr*(*b*) at stochastic time intervals characterized by $\varepsilon$, with exactly one copy of *null*(*b*) being preserved.

### 2.2 Gene Products

We now describe the transcription factor *tr*(*b*) (Figure 2), introducing the process calculus notions of interaction and stochastic choice. Except for delays $\tau$, which happen autonomously, any action that a process performs must happen in conjunction with a complementary action performed by another process. The simultaneous occurrence of complementary actions is an *interaction*, e.g. between two molecules, or between a transcription factor and a promoter site. An action can be *offered* at any time, but only complementarily offered actions can result in actual interactions. For an interaction site, or *channel*, *b*, such complementary actions are conventionally called *input on b* (written '?*b*'), and *output on b* (written '!*b*'). (In our examples we need only consider such simple *signaling* interactions; in general an interaction can also exchange data in the form of a message from output to input.) Hence, ?*b* and !*b* are complementary actions that can exchange a signal between them and allow two corresponding processes to change state.

The transcription factor *tr*(*b*) offers a choice of two actions; one is an output action !*b*, representing interaction with a binding site, and the other is a delay $\tau$, followed by

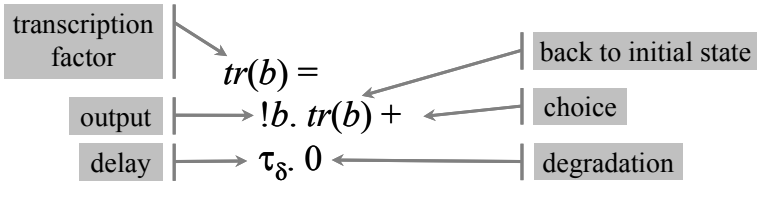



**Fig. 2.** A transcription factor *tr*(*b*) makes a *stochastic choice* ('+') between either binding to an available promoter site *b* by an *output action* ('!*b*'), or *delaying* ('$\tau$') with rate $\delta$. In the first case, the output action interacts with a corresponding input action at a promoter site *b*, and *then* ('.') the transcription factor returns to its initial state *tr*(*b*), ready to interact again. In the second case, the transcription factor degrades to the inert process ('0').

degradation. These two actions are in a stochastic race, indicated by '+': $b$ has (implicitly defined with it) a fixed associated rate $r$, and $\tau$ has a specific rate $\delta$. If $!b$ wins the race, it means that an interaction has occurred with an input action $?b$ offered elsewhere, and the process returns to the initial state, $tr(b)$. If $\tau$ wins the race, however, the following state is 0: the inert process that never performs any actions.

All together, $tr(b)$ is defined as $(!b.\ tr(b)) + (\tau_\delta.0)$, which means that $tr(b)$ has the potential to interact multiple times with promoter sites, but each time (and particularly if no promoter site is available) it has a chance to degrade. Without interactions with binding sites, a fixed population of transcription factors will simply exponentially degrade. If the population is being replenished, then a stable level may be found between production and degradation.

### 2.3 Unary Gates

We now consider gates with simple regulation. A $neg(a,b)$ gate has a promoter site $a$ with negative regulation (inhibition), and a product $b$.

The $neg(a,b)$ gate (Figure 3) has a subprocess that is essentially identical to the $null(b)$ gate, i.e., it provides constitutive transcription. However this subprocess is now in a stochastic race with a subprocess $?a.\ \tau_\eta.\ neg(a,b)$. That is, it is in a race with a promoter binding, $?a$. If the promoter component wins the race (by interacting with a transcription factor $tr(a)$), the + choice is taken on the promoter side, and the whole process becomes $\tau_\eta.\ neg(a,b)$. In this state, the gate is stuck performing a stochastic delay $\tau_\eta$, i.e., it is inhibited, after which it goes back to be $neg(a,b)$.

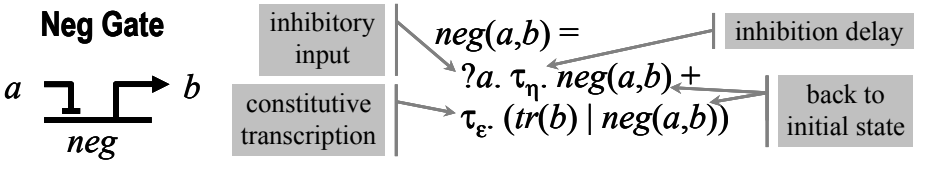



**Fig. 3.** A gene gate with inhibitory control, $neg(a,b)$ makes a stochastic choice ('+') between constitutive transcription and inhibitory stimulation. The constitutive transcription case (bottom line) is exactly as in Figure 1, but this time it is in a race with a stimulus. If an interaction happens with the input action '$?a$', then the gate enters a stochastic delay ('$\tau_\eta$'), during which the gate is inhibited, and then returns to the initial state.

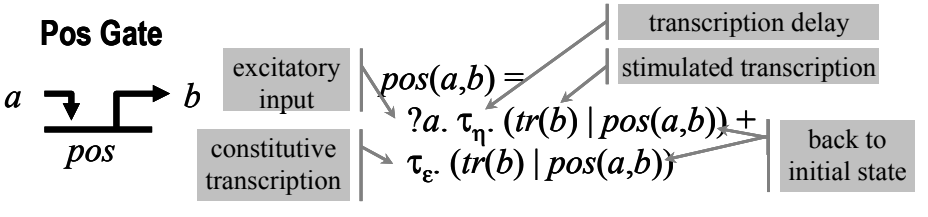



**Fig. 4.** A gene gate with excitatory control, $pos(a,b)$. This is almost identical to $neg(a,b)$, but the input stimulus '$?a$' is followed by the production of $tr(b)$ instead of an inhibitory delay.

The *pos*(*a*,*b*) gate (Figure 4) has a promoter site *a* with positive regulation (stimulation), and a product *b*. It is similar to the *neg* gate, but instead of an inhibition delay, we have a transcription delay followed by stimulated production of *tr*(*b*).

## 3 The Stochastic π-Calculus Execution Model

### 3.1 Simulation Language

We have seen how a biological system can be modeled in the stochastic π-calculus, by representing each component of the system as a process *P* that precisely describes what the component can do. To summarize, the most basic process form is a choice $\Sigma = P_1 + \ldots + P_n$ between zero or more outputs $!x(n)$, inputs $?x(m)$, and delays $\tau$ that the component can perform (in the general form of input/output, *n* is the output message and *m* is the input variable). Two components *P* and *Q* can be combined together using parallel composition *P*|*Q*. Channels can be established to allow the components to interact by complementary inputs and outputs. Once a biological system has been modeled using these basic components, the model can be stochastically simulated in order to predict the evolution of the system over time. In this paper, the simulations were obtained using the Stochastic Pi Machine (SPiM), which is described in [13].

Another basic operator of stochastic π-calculus, which we do not need to discuss in detail in this paper, allows the creation of fresh channels. The operator $new\ x_\varepsilon.\ P$ creates a fresh channel *x* of rate ε to be used in the process *P*. The rules of stochastic π-calculus ensure that a "fresh" channel so obtained does not conflict with any other channel. We mention the channel creation operator here just because it allows us to obtain the stochastic delay $\tau_\varepsilon$ as a derived operator. In fact, we can define:

$$\tau_\varepsilon.P + Q \quad = \quad new\ x_\varepsilon.\ (!x.0 \mid (?x.P + Q)) \qquad \text{for } x \text{ not occurring in } P \text{ or } Q$$

That is, a delay is equivalent to a single communication on a fresh channel of the same rate. Hence, stochastic delays can be reduced to ordinary channel communication, and can be handled uniformly like any other communication, e.g., for simulation purposes.

### 3.2 Simulator

The Stochastic Pi Machine simulates a given process *P* by first converting the process to a corresponding simulator data structure, consisting of a list of components $A = \Sigma_1, \ldots, \Sigma_M$. The resulting list is then processed by the simulator, by first using a function *Gillespie*(*A*) to stochastically determine the next interaction channel *x* and the corresponding reaction time $\tau$. Once an interaction channel *x* has been chosen, the simulator uses a *selection operator* to randomly select from the list *A* a component of the form $\Sigma + ?x(m).P$ containing an input on channel *x*, and different component of the form $\Sigma' + !x(n).Q$ containing an output on *x*. The selected components can then interact by synchronizing on channel *x*, and the processes *P* (with the input variable *m* replaced by *n*) and *Q* are added to the remainder of the list. The simulator continues processing the list in this way until no more interactions are possible.

The function *Gillespie*(*A*) is based on [14], which uses a notion of *channel activity* to stochastically choose a reaction channel from a set of possible channels. The activity of a reaction channel corresponds to the number of possible combinations of reactants on the channel; channels with a high activity and a fast reaction rate have a higher probability of being selected. A similar notion of activity is defined for the Stochastic Pi Machine, where $Act_x(A)$ denotes the number of possible combinations of inputs and outputs on interaction channel $x$ in a list of components $A$:

$$Act_x(A)=(In_x(A)*Out_x(A))-Mix_x(A)$$

$In_x(A)$ and $Out_x(A)$ are defined as the number of available inputs and outputs on interaction channel $x$ in $A$, respectively, and $Mix_x(A)$ is the sum of $In_x(\Sigma_i)\times Out_x(\Sigma_i)$ for each component $\Sigma_i$ in $A$. The formula takes into account the fact that an input and an output in the same component cannot interact, by subtracting $Mix_x(A)$ from the product of the number of inputs and outputs on $x$.

The Stochastic Pi Machine has been formally specified in [13], and the specification has been proved to correctly simulate π-calculus processes. The simulator has also been used to simulate a wide variety of chemical and biological systems. In particular, many of the benchmark examples that were used to validate the Gillespie algorithm [14] have been modeled as π-calculus processes and correctly simulated in SPiM.

### 3.3 Interaction-Oriented Simulation vs. Reaction-Oriented Simulation

The Gillespie algorithm was originally used to simulate a set of chemical reaction equations expressed in terms of reactants and products, and the results of a simulation were plotted as the quantity of each chemical species versus time. In contrast, the π-calculus does not describe an equation for each type of chemical reaction, but instead describes the behavior of each component in terms of the inputs and outputs it can perform on a set of interaction channels. This gives rise to an interaction-oriented model, as opposed to a chemical-reaction-oriented model, in which a reactant is defined as an input or output on a given interaction channel. Once the notion of a reactant has been defined in this way, the Gillespie algorithm can be directly applied to a given π-calculus model of the biological system. The corresponding simulation results can be plotted as the quantity of each reactant versus time.

## 4 Gene Networks

### 4.1 Simple Circuits

In Section 2 we have described gene gates with one input; gates with *n* inputs can be defined similarly, to form a larger library of components. Once the components are defined, gene circuits can be assembled by providing interaction channels, with associated interaction rates, connecting the various gates. If we write, e.g., *pos*(*a*,*b*) | *neg*(*b*,*c*), the *pos* process will offer output actions !*b*, through *tr*(*b*), and the *neg* process will offer input actions ?*b*. Hence the shared channel *b*, given to both *pos* and *neg* as a parameter, can result in repeated interactions between the two processes over *b*, and hence in network connectivity.

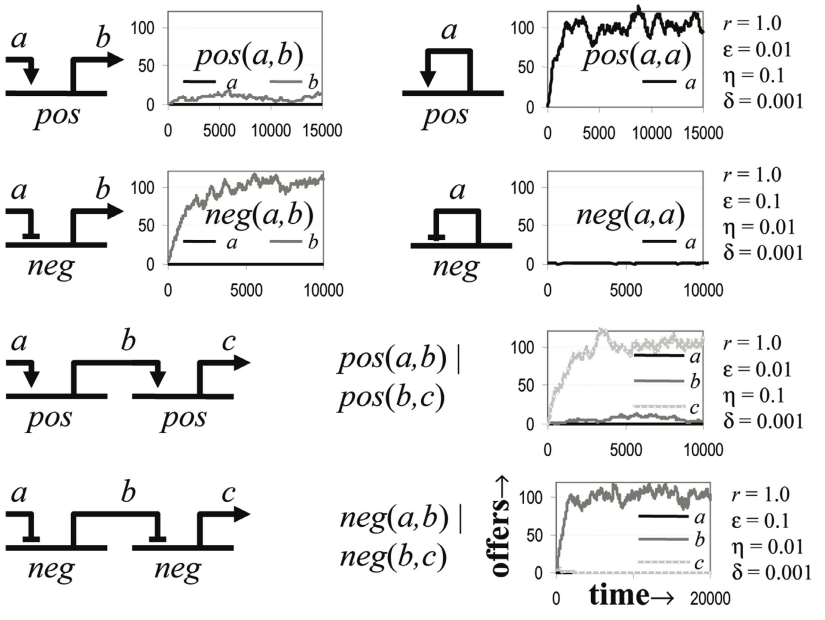



**Fig. 5.** Compositions of gates represent circuits (left) that exhibit behaviors (right). The channels *a,b,c* are declared separately (not shown) along with their associated stochastic interaction rates. In all simulations, the common rate *r* for *a,b,c* is set to a baseline value of 1.0. The other chosen rates are as indicated in the individual simulations; the fact that they are chosen at simple order-of-magnitude intervals suggests that they are not critical for the intended behavior. The vertical axis is the number of outstanding *offers* of communication: for a channel *a* we may plot output offers !*a* or input offers ?*a*. In all cases above, the networks get started by constitutive transcription only. All the plots are of individual simulator runs.

The simplest circuits we can build are single gates interacting with themselves in a feedback loop, like *pos*(*a*,*a*) (Figure 5). In absence of any stimulus on *a*, *pos*(*a*,*a*), must choose the constitutive transcription route and evolve into *tr*(*a*) | *pos*(*a*,*a*), where now *tr*(*a*) can stimulate *pos*(*a*,*a*) at a faster rate than the constitutive rate, and possibly multiple times. Depending on the production and degradation rates, a stable high level of *tr*(*a*) may be reached. Similarly *neg*(*a*,*a*) can stabilize at a low quantity of *tr*(*a*) where degradation of *tr*(*a*) balances inhibition. A convenient high-signal level of about 100 is maintained in our examples by appropriate rates (see parameters in Figure 5).

The combination *pos*(*b*,*a*) | *neg*(*a*,*b*) (Figure 6) is a self-inhibition circuit, like *neg*(*a*,*a*), and it similarly has a stable output. But now there are two separate products, *tr*(*a*) and *tr*(*b*), so the system (again in absence of any stimulus) can stochastically start with a prevalence of *tr*(*a*) or a prevalence of *tr*(*b*): this can be seen at the beginning of the two plots, before stabilization.

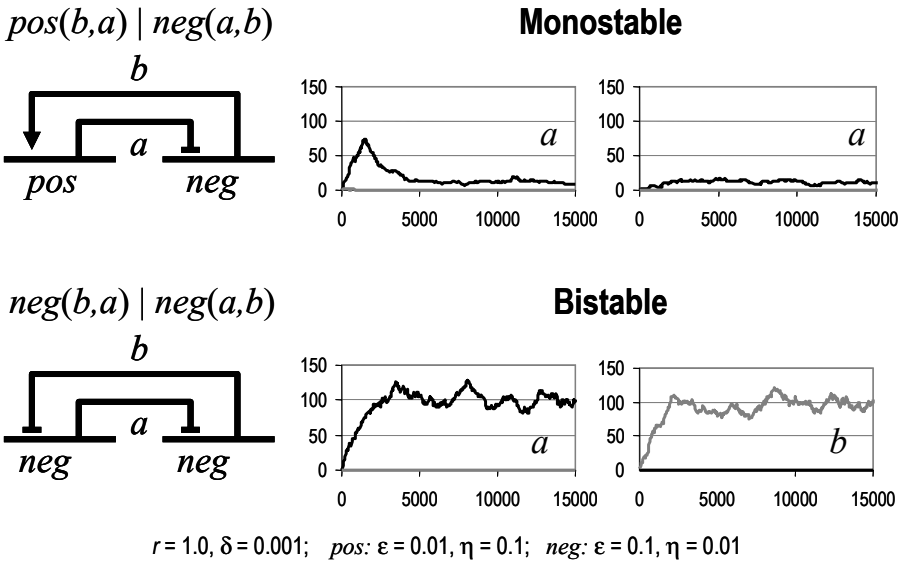



**Fig. 6.** Feedback loops that are *monostable* (resulting in a single stable state with *a* high after a transient) and *bistable* (resulting in two distinct stable states with *a* high or *b* high)

The combination *neg*(*b*,*a*) | *neg*(*a*,*b*) (Figure 6) is a bistable circuit, which can start up in one state or another, and (usually) stay there.

### 4.2 Repressilator

The well-known repressilator circuit [12], consisting of three *neg* gates in a loop, is an oscillator. We compare here three different degradation models, aiming to justify somewhat our initial definition for *tr*(-). In the first model (Figure 7(A)), each transcription factor interacts exactly once, and only then it disappears. The repressilator circuit oscillates nicely but, without stochastic degradation, the plots appear very "mechanical"; moreover, the quantities of products grow at each cycle because products do not disappear unless they interact. In the second model (Figure 7(B)), each transcription factor interacts exactly once, or can degrade. Again the plots look mechanical, but the stochastic degradation defines a stable level of product. The third model (Figure 7(C)), with multiple interactions and stochastic degradation, is more realistic and gives more convincing plots. See the Appendix for the simulator script.

The progressive refinement of the definition of *tr*(-), provides an illustration of how one can play with process descriptions to find models that show a balance between simplicity and realism. A further step could be to model both attachment and detachment of transcription factors, and then to model both transcription and translation.

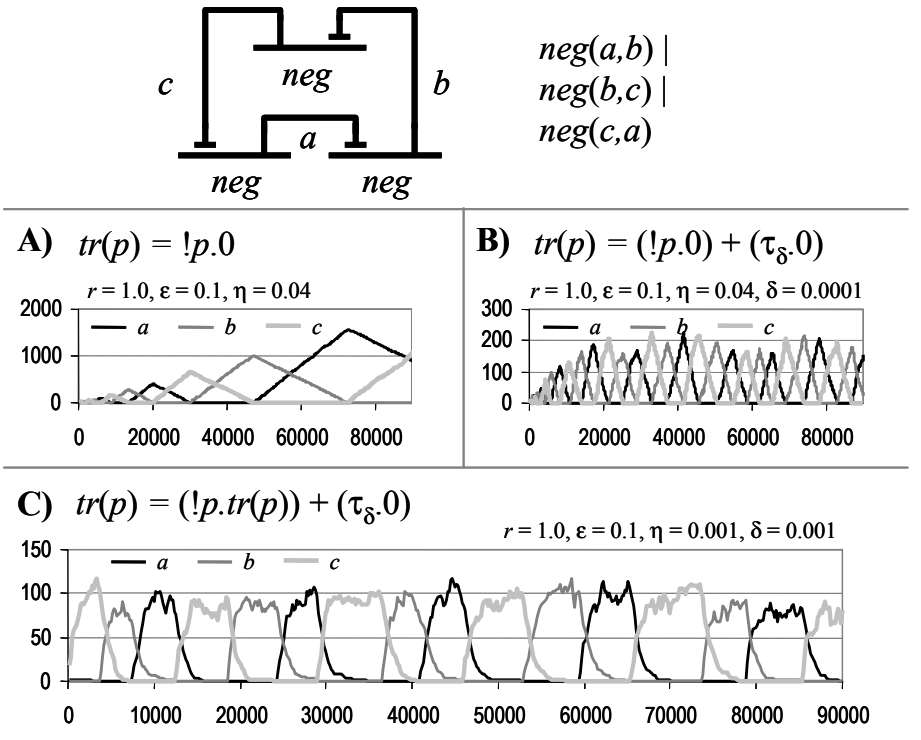



**Fig. 7.** The Repressilator circuit and its dynamics for different degradation models (A – C). The detailed explanation is found in the text.

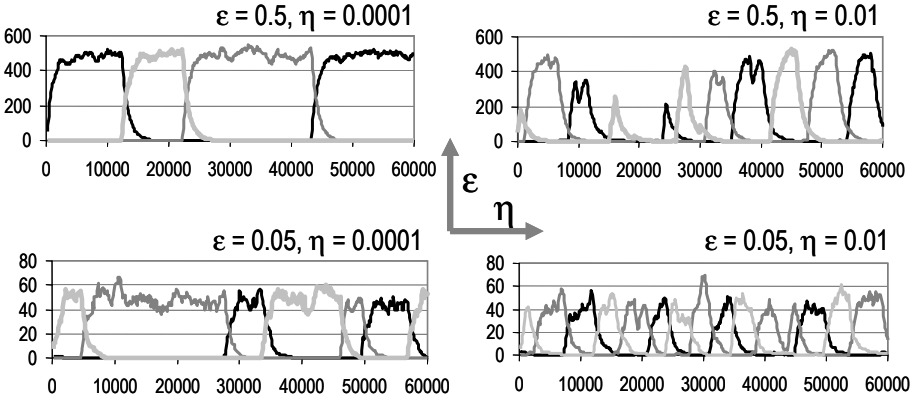



**Fig. 8.** Repressilator frequency and amplitude, regulated by η and ε. *Cf.* Figure 7(C).

### 4.3 Network Properties: Oscillation

It is instructive to take a "systems" approach and see what the rate parameters described earlier mean in the context of networks of gates. In the case of the

repressilator we can see that the constitutive rate (together with the degradation rate) determines oscillation amplitude, while the inhibition rate determines oscillation frequency. Figure 8 shows the variation of ε and η from their values in Figure 7(C)); note the differences in scale.

Moreover, we can view the interaction rate *r* as a measure of the volume (or temperature) of the solution; that is, of how often transcription factors bump into gates. Figure 9 shows that the oscillation frequency and amplitude remain unaffected in a large range of variation of *r* from its value in Figure 7(C)). Note that *r* is in a stochastic race against δ in *tr*, and δ is always much slower.

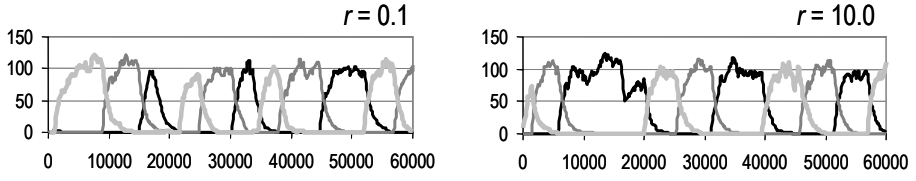



**Fig. 9.** Repressilator stability to changes in *r* (volume/temperature). *Cf.* Figure 7(C).

### 4.4 Network Properties: Fixpoint

We now discuss a network property that becomes important in later analysis. Figure 10 plots signals flowing through a sequence of *neg* gates with parameters as in Figure 7(C), except for η, the inhibition delay. On the left, the signals are alternating between high (*b*,*d*) and low (*a*,*c*,*e*). As η is increased, shown from left to right, the gates behave less and less like boolean operators, but the signals remain separate.

Figure 11 shows the same circuit, except for a self feedback on the head gate. With low inhibition delay η (i.e. ineffective feedback) the system is unstable (left). But soon after, as we increase η, the self feedback flattens *all* signals downstream to a common low level (middle). The signals remain at a common level over a wide range of η, although this level is raised by increasing η (right).

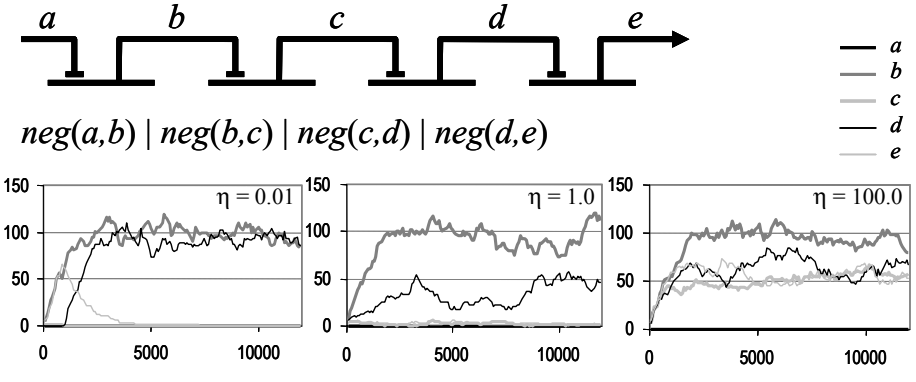



**Fig. 10.** A sequence of *neg* gates with three settings of their η parameter

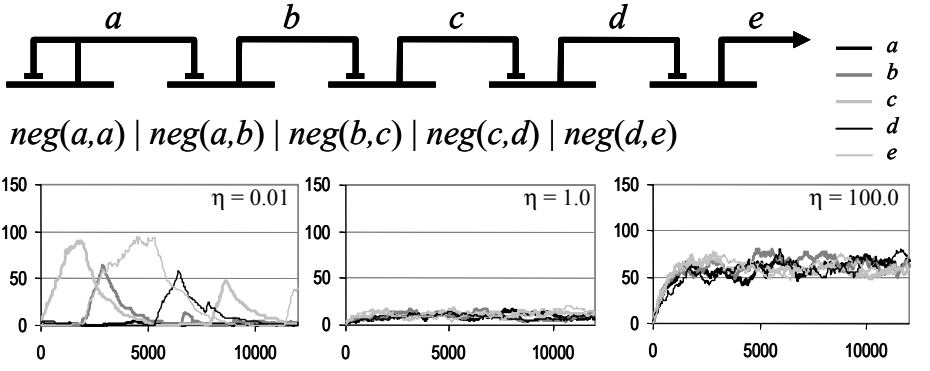



**Fig. 11.** The effect of head feedback on a sequence of *neg* gates

This behavior is self-regulating, and can be explained as follows. The head feedback naturally finds a fixpoint where gate input equals gate output (unless it oscillates). If the next gate has the same parameters, its output will then also equal its input, and so on down the line: all the gates will be at the same fixpoint. Different values of η and different gate response profiles may change the fixpoint level, but not its fundamental stability.

### 4.5 Combinatorial Circuits

As examples of non-trivial combinatorial networks and their stochastic simulation, we now examine the artificial gene circuits described by Guet *et al.* [1]. Most of those circuits are simple combinations of inhibitory gates exhibiting expected behavior. However, it was found that in some of the circuits subtle (and partially still not understood) behavior arises; we focus particularly on two of these cases.

In order to build up the different combinatorial networks easily, we begin with a version of the *neg* gate that is more flexibly parameterizable. We call it *negp*, and it has the property that, if *s* represents the rates used in the *neg* gate, then $negp(a,s,tr(b)) = neg(a,b)$, hence *neg* is a special case of *negp*. The rates for inhibition and constitutive translation are passed as a pair $s=(\varepsilon,\eta)$, in the second parameter. The third parameter fully encapsulates the gate product, so the gate logic is independent of it[1].

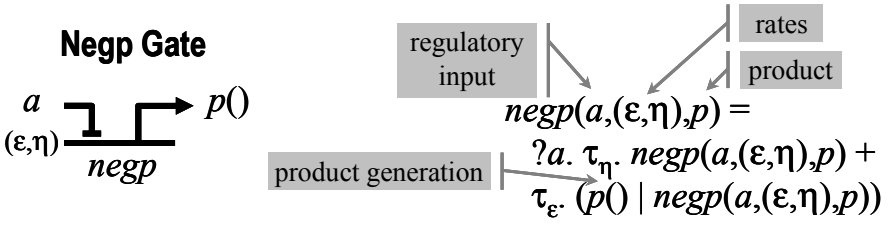



**Fig. 12.** A *neg* gate with parametric product *p*

[1] More technically, if we set *pb*() = *tr*(*b*) (*pb* is the process that when invoked with no arguments, invokes *tr* with argument *b*), then we have *negp*(*a*,*s*,*pb*) = *neg*(*a*,*b*); we write *negp*(*a*,*s*,*tr*(*b*)) as an abbreviation, skipping the intermediate definition of *pb*.

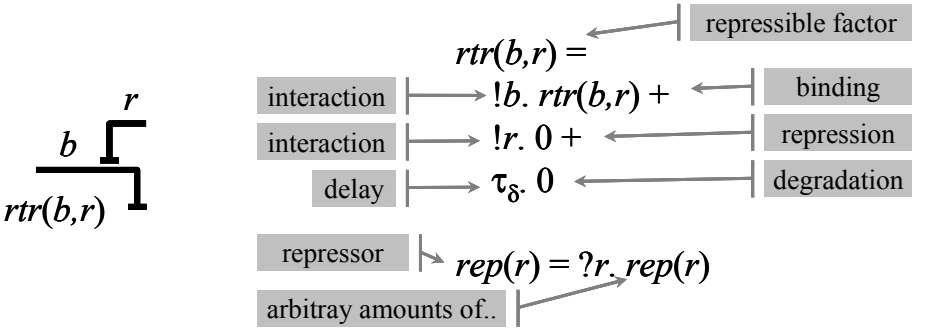



**Fig. 13.** Repressible transcription factors

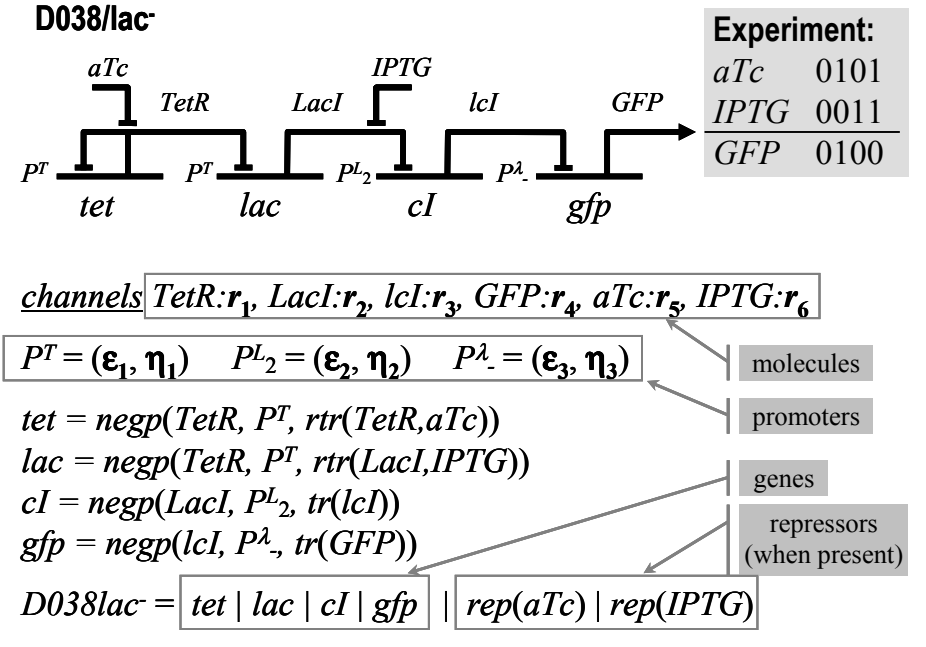


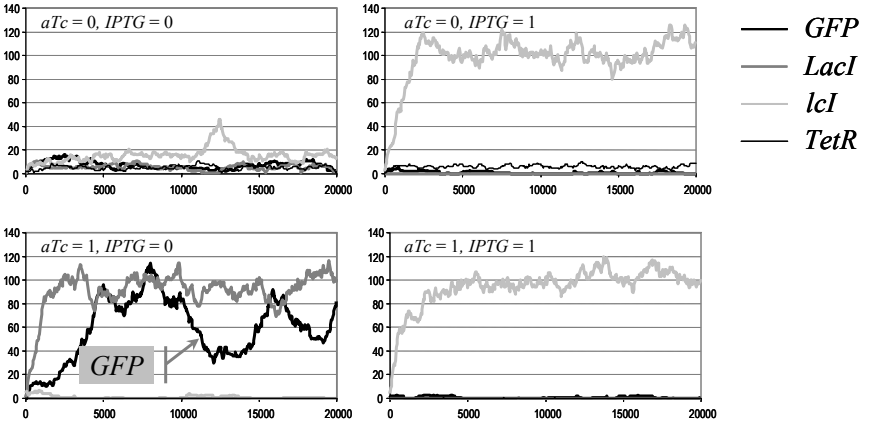

**Fig. 14.** D038

In addition to the old transcription factors *tr*(*b*), binding to a site *b*, we now need also transcription factors that can be repressed: *rtr*(*b*,*r*). These have three possible behaviors: binding to a site *b*, being neutralized via a site *r*, and degrading. The repression is performed by a process *rep*(*r*) that, if present, "inexhaustibly" offers ?*r*.

In the artificial gene circuits by Guet et al, the circuits are probed by varying two inputs: two so-called "inducer" proteins in the environment, aTc and IPTG, which bind specifically to the gene in question. The output of the gene circuit is detected by a reporter gene which produces a green-fluorescent protein (*GFP*) which can be optically detected.

We can now describe the circuits from [1] by simple combinations of *negp*, *tr*, *rtr*, and *rep* components. All the other names appearing here, such as *TetR*, *aTc*, etc., which glue the network together, are just channel names used in complementary input and output actions.

Intuitive Boolean analysis of one of the still controversial circuits, D038, in Figure 14 would suggest either oscillation (*GFP*=0.5 on average), or *GFP*=1, contrary to experiment[2]. Thus, for the given construction, a different explanation is needed. The fixpoint effect, however, which we have described in Section 4.4, does suggest an explanation for the output in the absence of repressors, whereby all signals including the output signal *GFP* are driven to a fixpoint with a low value. The addition of *GFP* renders that state unstable and drives *TetR* to 0, and hence *GFP* to 1. In all cases, the addition of *IPTG* drives *LacI* to 0 and hence *GFP* to 0. Figure 15 shows the simulation results of this system for the different values of *aTc* and *IPTG*. In circuit D038 we have thus found an example in which the modelling of the stochastic gate behaviour can indeed help to find an explanation of the observed dynamics.




$r_{1..6} = 1.0,\ \delta = 0.001$
$\varepsilon_{1,2,3} = 0.1,\ \eta_1 = 0.25\ (P^T),\ \eta_{2,3} = 1.0\ (P^L{}_2,\ P^\lambda{}_-)$

**Fig. 15.** D038 simulations

[2] In absence of repressors, the experimentally observed *GFP* is 0 (meaning no detectable signal), hence, by tracing boolean gates backwards, *lcI*=1, and *LacI*=0, and *TetR*=1. But by self-loop *TetR*=1 implies *TerR*=0, so the whole circuit, including *GFP* should be oscillating and averaging *GFP*=0.5. As an alternative analysis, consider the level of *TetR* (which is difficult to predict because it is the result of a negative self-feedback loop). Whatever that level is, and whether or not *aTc* is present, it must equally influence the *tet* and *lac* genes, since the promoters are the same ($P^T$). The option, *TetR*=*LacI*=1 gives *GFP*=1. Suppose instead *TetR*=*LacI*=0, then *lcI*=1, and *GFP*=0 as observed. But in that situation, with *TetR*=0, *aTc* should have no influence, since it can only reduce the level of *TetR*. Instead, *aTc* somehow pushes *GFP* to 1.

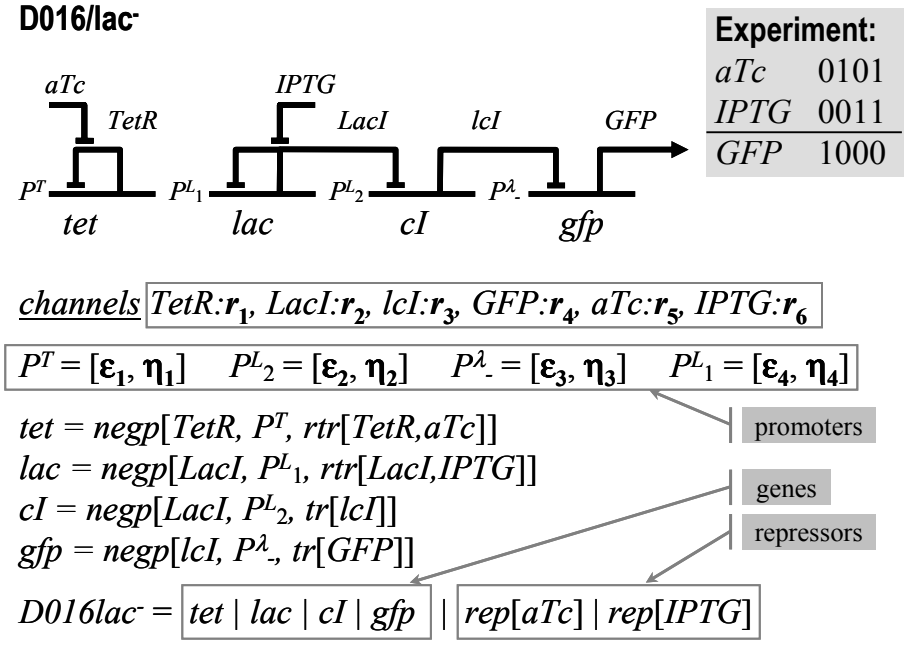



**Fig. 16.** D016

In a very similar fashion we can code another peculiar circuit, D016, shown in Figure 16. This circuit is perplexing because addition of *aTc*, affecting an apparently disconnected part of the circuit, changes the *GFP* output. In [18] it is suggested that this may be caused by an overloading of the degradation machinery, due to an overproduction of *TetR* when *aTc* is present, which might decrease the degradation rate of the other proteins. But even in absence of *aTc* and *IPTG*, it is surprising that *GFP* is high (about 50% of max [16]): this seems to contradict both simple boolean analysis and our fixpoint explanation which worked well for D038.

One way to rationalize the behaviour displayed by this circuit is to assume that the $P^L{}_1$-*lac* gate is operating in a region in parameter space in which the circuit dynamics is unstable. A closer examination of the instability region of our basic fixpoint circuit (Figure 10 bottom left) shows that, while the first signals in the sequence (*a*,*b*) are kept low, the subsequent signals (*c*, corresponding to *GFP* in D016, and *d*,*e*) all spike frequently. This may give the appearance, on the average, of high levels of *GFP*, matching the first column of the D016 experiment. Moreover, in the instability region the system responds very sensitively to changes in degradation levels: *GFP* levels can be brought down both by increasing degradation by a factor of 5 (because this brings the circuit back into the fixpoint regime) or by decreasing degradation by a factor of 1000 (so that there are enough transcription factors to inhibit all gates). In Figure 17 we begin by placing D016 in the instability regime, with *GFP* spiking (A). Then, adding *aTc* while reducing degradation suppresses all signals (B). Adding *IPTG* results in no *GFP* (C,D); moreover, reduced degradation causes overproduction (D). Even increased degradation (E) can result in no *GFP*.

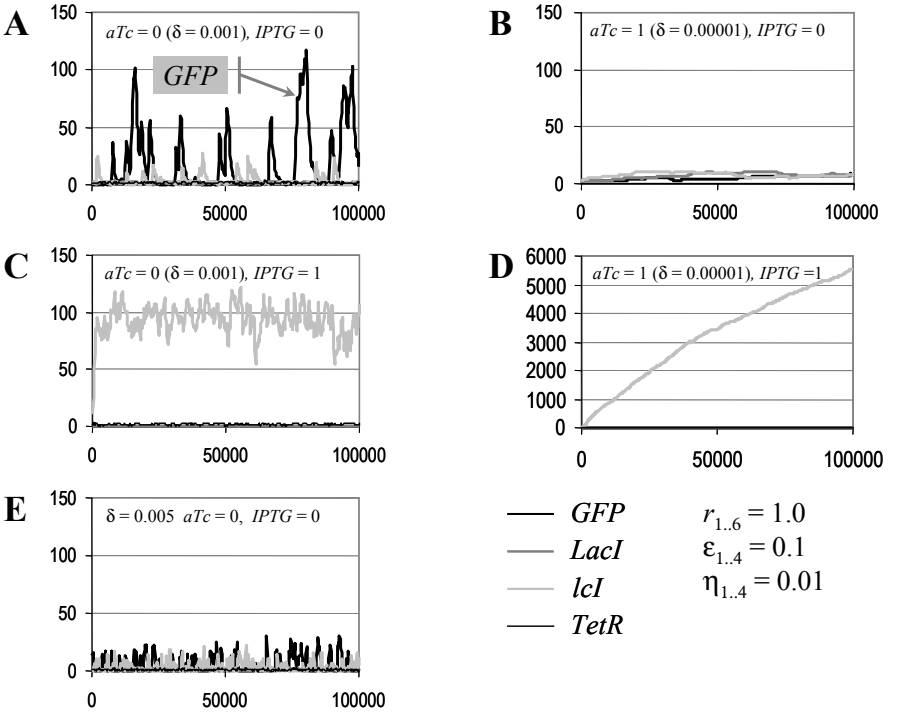



**Fig. 17.** D016 simulations

While a proper biological explanation of the behavior of D016 has not been obtained yet, the type of analysis we have performed here already shows the potential of the information gain from a proper study of the stochastic dynamics of the gene circuits, in particular in the case where head feedbacks are present; other authors have noted the possibility of surprises in such cases [19].

## 5 Conclusions

In this paper we have demonstrated how stochastic simulations of gene circuits can be built in a compositional way by employing the stochastic π-calculus. For this, we chose as a descriptive level not the molecular constituents, but rather considered each gene as a gate with corresponding inputs and outputs. On this level, compositionality is illustrated, for example, by our treatment of the repressilator circuit: the definition of the *neg* gate could be left unchanged when the definition of the transcription factor *tr* was refined. Our approach is mechanistic in the sense that we (re-)construct a biological system from discrete elements and then deduce the system behaviour as arising from the interactions of the components. This differs from modelling attempts of the same systems in the bioinformatics literature which only looked at gene expression levels without considering their origin [18]. Our approach, while being abstract, is advantageous as it allows a considerable flexibility in the level of detail

with which components and their interactions are described (see the Appendix for further illustration). While the adopted level of the description may be considered coarse and qualitative, the π-calculus approach easily allows for refinements (i.e., inclusion of additional detail down to molecular levels of description) to match available knowledge.

Apart from these analytical and conceptual advantages in building up the different circuits, we stress that the ease of use of the compositional approach in combination with stochastic simulations is particularly useful for hypothesis testing. It can build on available knowledge, but the outcome of the stochastic simulations of the interacting components yields a highly non-trivial check of expectations. By comparison, Boolean analysis or intuitive ideas are obviously too naïve and thus can easily be misleading.

The sensitivity of the gene network dynamics to parameter choice has to be contrasted with the lack of quantitative knowledge of promoter strengths, or even qualitative relationships between the different promoters [19]. In the absence of "true" (i.e., experimentally validated) parameter values, a detailed analysis of the stochastic behaviour of the gene networks resulting from a systematic parameter variation can be a very useful - but clearly not sufficient - step to avoid misinterpretations of experiments.

To conclude, we believe that the compositional approach we propose for the formulation of stochastic models of gene networks will allow a useful path for more detailed, quantitative studies of regulatory mechanisms, and in particular for the testing of hypotheses of complex system behavior. It may be considered as one step towards the development of flexible languages and simulation tools for computational biology, for which a need has recently been expressed by several biologists ([20]-[22]).

## References

[1] Guet, C.C., Elowitz, M.B., Hsing, W. & Leibler, S. (2002) Combinatorial synthesis of genetic networks. Science 296 1466-1470.

[2] Thattai, M. & van Oudenaarden, A. (2001) Intrinsic noise in gene regulatory networks. Proc. Nat. Acad. Sci. 98, 8614- 8619.

[3] Paulsson, J., Berg, O.G. & Ehrenberg M. (2000) Stochastic Focusing: fluctuation-enhanced sensitivity of intracellular regulation. Proc. Nat. Acad. Sci. 97, 7148-7153.

[4] Milner, R. (1999) Communicating and Mobile Systems: The π-Calculus. Cambridge University Press.

[5] Priami, C., Regev, A., Shapiro, E. & Silverman, W. (2001) Application of stochastic process algebras to bioinformatics of molecular processes. Information Processing Letters 80 25-31.

[6] Regev, A. (2002) Computational Systems Biology: A Calculus for Biomolecular knowledge. Ph.D. Thesis, Tel Aviv University.

[7] Regev, A. & Shapiro, E. (2002) Cellular abstractions: Cells as computation. Nature 419 343.

[8] Regev, A., Panina, E.M., Silverman, W., Cardelli, L. & Shapiro, E. (2004) BioAmbients: An abstraction for biological compartments. Theoretical Computer Science, 325(1) 141-167.

[9] Cardelli, L. (2004) Brane Calculi - Interactions of Biological Membranes. Computational Methods in Systems Biology. Springer. 257-278.
[10] Chiarugi, D., Curti, M., Degano, P. & Marangoni, R.: VICE: A VIrtual CEll. CMSB 2004: 207-220.
[11] Kuttler, C.& Niehren, J. (2005) Gene Regulation in the Pi Calculus: Simulating Cooperativity at the Lambda Switch. Transactions on Computational Systems Biology, to appear.
[12] Elowitz, M.B., Leibler. S. (2000) A synthetic oscillatory network of transcriptional regulators. Nature 403 335-338.
[13] Philips, A. & Cardelli, L., (2005). A Correct Abstract Machine for the Stochastic Pi-calculus. Proc. BioConcur 2004.
[14] Gillespie, D. (1977) Exact stochastic simulation of coupled chemical reactions. J. Chem. Phys. 81 2340-2361.
[15] Hillston, J.: A Compositional Approach to Performance Modelling. Cambridge University Press, 1996.
[16] Guet, C.C., personal communication.
[17] Wigler, M. & Mishra, B. (2002) Wild by nature. Science 296 1407-1408.
[18] Mao, L. & Resat, H. (2004) Probabilistic representation of gene regulatory networks. Bioinformatics 20 2258-2269.
[19] Ronen, M., Rosenberg, R., Shraiman, B.I. & Alon, U. (2002) Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. Proc. Nat. Acad. Sci. 99 10555-10560.
[20] Brenner, S. (1995) Loose Ends. Curr. Biology 5 332.
[21] Bray, D. (2001) Reasoning for Results. Nature 412 863.
[22] Lazebnik, Y. (2002) Can a biologist fix a radio? Or, what I learned while studying apoptosis. Cancer Cell 2 179-182.

## Appendix

### A Simulator for the Stochastic π-Calculus

The following is a detailed description of the Stochastic π-calculus and the Stochastic Pi Machine, as presented in [13].

| | | | | | |
|---|---|---|---|---|---|
| $P,Q$::= | new $x$ $P$ | Restriction | $\Sigma$::= | $\mathbf{0}$ | Null |
| \| | $P \mid Q$ | Parallel | \| | $\pi.P + \Sigma$ | Action |
| \| | $\Sigma$ | Choice | $\pi$::= | $!x(n)$ | Output |
| \| | $*\pi.P$ | Replication | \| | $?x(m)$ | Input |

**Def. 1.** Syntax of the Stochastic π-calculus

$$!x(n).P + \Sigma \mid ?x(m).Q + \Sigma' \xrightarrow{rate(x)} P \mid Q_{\{n/m\}} \quad [1]$$

$$P \xrightarrow{r} P' \Rightarrow P \mid Q \xrightarrow{r} P' \mid Q \quad [2]$$

$$P \xrightarrow{r} P' \Rightarrow \text{new } x\ P \xrightarrow{r} \text{new } x\ P' \qquad [3]$$

$$Q \equiv P \xrightarrow{r} P' \equiv Q' \Rightarrow Q \xrightarrow{r} Q' \qquad [4]$$

**Def. 2.** Reduction in the Stochastic π-calculus

---

**Stochastic π-calculus.** A biological system can be modeled in the stochastic π-calculus by representing each component of the system as a calculus process $P$ that precisely describes what the component can do. According to Def. 1, the most basic component is a choice Σ between zero or more output $!x(n)$ or input $?x(m)$ actions that the component can perform. Two components $P$ and $Q$ can be combined together using parallel composition $P|Q$, and a component $P$ can be given a private interaction channel $x$ using restriction *new x P*. In addition, multiple copies of a given component $\pi.P$ can be cloned using replication $*\pi.P$. Standard syntax abbreviations are used, such as writing π for $\pi.\mathbf{0}$ and $\pi.P$ for $\pi.P + \mathbf{0}$.

Two components in a biological system can interact by performing complementary input and output actions on a common channel. During such an interaction, the two components can also exchange information by communicating values over the channel. Each channel $x$ is associated with a corresponding interaction rate given by *rate*($x$) and the interaction between components is defined using reduction rules of the form $P \xrightarrow{r} P'$. Each rule of this form describes how a process $P$ can evolve to $P'$ by performing an interaction with rate $r$. According to Def. 2, a choice containing an output $!x(n).P$ can interact with a parallel choice containing an input $?x(m).Q$. The interaction occurs with *rate*($x$), after which the value $n$ is assigned to $m$ in process $Q$ (written $Q_{\{n/m\}}$) and processes $P$ and $Q_{\{n/m\}}$ are executed in parallel (Eq. 1). Components can also interact in parallel with other components (Eq. 2) or inside the scope of a private channel (Eq. 3), and interactions can occur up to re-ordering of components (Eq. 4), where $P \equiv Q$ means that the component $P$ can be re-ordered to match the component $Q$. In particular, the re-ordering $*\pi.P \equiv \pi.(P \mid *\pi.P)$ allows a replicated input $*?x(m).Q$ to clone a new copy of $Q$ by reacting with an output $!x(n).P$.

---

| | | | | | |
|---|---|---|---|---|---|
| $V,U$::= | new $x$ $V$ | Restriction | $A,B$::= | [] | Empty |
| | $A$ | List | | Σ:: $A$ | Choice |

**Def. 3.** Syntax of the Stochastic Pi Machine

---

$$\begin{array}{l} x,\tau = \mathit{Gillespie}(A) \\ \wedge\ A > (?x(m).P + \Sigma) :: A' \\ \wedge\ A' > (!x(n).Q + \Sigma') :: A'' \end{array} \Rightarrow A \xrightarrow{rate(x)} P_{\{n/m\}} : Q : A'' \qquad [5]$$

$$V \xrightarrow{r} V' \Rightarrow \text{new } x\ V \xrightarrow{r} \text{new } x\ V' \qquad [6]$$

**Def. 4.** Reduction in the Stochastic Pi Machine

---

**Stochastic Pi Machine.** The Stochastic Pi Machine is a formal description of how a process of the stochastic π-calculus can be simulated. A given process *P* is simulated by first encoding the process to a corresponding simulator term *V*, consisting of a list of choices with a number of private channels:

$$\text{new } x_1 \ldots \text{new } x_N \ \ (\Sigma_1::\Sigma_2::\ldots::\Sigma_M::[])$$

This term is then simulated in steps, according to the reduction rules in Def. 4. A list of choices *A* is simulated by first using a function *Gillespie*(*A*) to stochastically determine the next interaction channel *x* and the corresponding interaction time τ. Once an interaction channel *x* has been chosen, the simulator uses a *selection operator* (>) to randomly select a choice $?x(m).P + \Sigma$ containing an input on channel *x* and a second choice $!x(n).Q + \Sigma'$ containing an output on *x*. The selected components can then interact by synchronizing on channel *x*, where the value *n* is sent over channel *x* and assigned to *m* in process *P* (written $P_{\{n/m\}}$). After the interaction, the unused choices Σ and Σ' are discarded and the processes $P_{\{n/m\}}$ and *Q* are added to the remainder of the list to be simulated, using a construction operator (:) (Eq. 5). An interaction can also occur inside the scope of a private channel (Eq. 6). The simulator continues performing interactions in this way until no more interactions are possible.

The function *Gillespie*(*A*) is based on the Gillespie Algorithm [14], which uses a notion of *channel activity* to stochastically choose a reaction channel from a set of available channels. The activity of a channel corresponds to the number of possible combinations of reactants on the channel. Channels with a high activity and a fast reaction rate have a higher probability of being selected. A similar notion of activity is defined for the Stochastic Pi Machine, where $Act_x(A)$ denotes the number of possible combinations of inputs and outputs on channel *x* in *A*:

$$Act_x(A) = In_x(A) \times Out_x(A) \ - \ Mix_x(A)$$

$In_x(A)$ and $Out_x(A)$ are defined as the number of available inputs and outputs on channel *x* in *A*, respectively, and $Mix_x(A)$ is the sum of $In_x(\Sigma_i)\times Out_x(\Sigma_i)$ for each choice $\Sigma_i$ in *A*. The formula takes into account the fact that an input and an output in the same choice cannot interact, by subtracting $Mix_x(A)$ from the product of the number of inputs and outputs on *x*. Once the values *x* and τ have been calculated, the simulator increments the simulation time by delay τ and uses the selection operator to randomly choose one of the available interactions on *x* according to (Eq. 5). This is achieved by randomly choosing a number $n\in[1..In_x(A)]$ and selecting the *n*th input in *A*, followed by randomly selecting an output from the remaining list in a similar fashion. The application of the Gillespie algorithm to the Stochastic Pi Machine is summarized in Def. 3, where *fn*(*A*) denotes the set of all channels in *A*.

---

1. For all $x\in fn(A)$ calculate $a_x = Act_x(A) \times rate(x)$
2. Store non-zero values of $a_x$ in a list $(x_\mu,a_\mu)$, where $\mu\in 1...M$.
3. Calculate $a_0=\sum_{\nu=0}^{M} a_\nu$

4. Generate two random numbers $n_1, n_2 \in [0,1]$ and calculate $\tau, \mu$ such that:

$$\tau = (1/a_0)\ln(1/n_1)$$

$$\sum_{\nu=1}^{\mu-1} a_\nu < n_2 a_0 \leq \sum_{\nu=1}^{\mu} a_\nu$$

5. *Gillespie*$(A) = (x_\mu, \tau)$.

**Def. 5.** Calculating *Gillespie(A)* according to (13)

---

For improved efficiency, the simulator can be modified to store a list of values for each channel $x$ in $A$, of the form:

$$x, In_x(A), Out_x(A), Mix_x(A), a_x$$

After each reduction has been performed, it is only necessary to update the values for those channels that were affected by the reduction, and then use Def. 5 on the updated values to choose the next reaction channel and calculate the delay.

To gain confidence in our simulation technique, we have conducted detailed simulations of the model chemical systems which were simulated in [14] using the Gillespie algorithm. Comparable results were obtained by modeling each system as a $\pi$-calculus process and simulating the resulting processes in the Stochastic Pi Machine.

**Repressilator Code**

From the simple examples discussed previously, the structure of the SPiM programs should now be clear. The following is the complete code for the repressilator simulation in Figure 7(C) of the paper, for the SPiM simulator (v0.04). In order to clarify parts of the code, comments are added in (* … *) brackets.

```
(* Simulation time, samples, and plotting  *)
directive sample 90000.0 500
directive plot !a as "a"; !b as "b"; !c as "c"

(* Parameters *)
val dk = 0.001           (* Decay rate *)
val inh = 0.001          (* Inhibition rate *)
val cst = 0.1            (* Constitutive rate *)
val bnd = 1.0            (* Protein binding rate *)

(* Transcription factor  *)
let tr(p:chan()) =
  do !p; tr(p)
  or delay@dk
```

```
(* Neg gate *)
let neg(a:chan(), b:chan()) =
  do ?a; delay@inh; neg(a,b)
  or delay@cst; (tr(b) | neg(a,b))

(* The circuit *)
new a @ bnd: chan()
new b @ bnd: chan()
new c @ bnd: chan()

run (neg(c,a) | neg(a,b) | neg(b,c))
```

**D038,D016 Code**

The following is the complete code for the of the D038 and D016 simulations in Figure 15 and Figure 17, for the SPiM simulator (v0.04).

```
(* Simulation time, samples, and plotting  *)
directive sample 20000.0 500
directive plot !GFP as "GFP"; !LacI as "LacI";
          !LambcI as "LambcI"; !TetR as "TetR"

(* Degradation rate *)
val dk = 0.001
(* val dk = 0.00001    for D016 when aTc is present *)

(* Transcription factor *)
let tr(b:chan()) =
  do !b; tr(b)
  or delay@dk

(* Repressible transcription factor *)
let rtr(b:chan(), r:chan()) =
  do !b; rtr(b,r)
  or !r
  or delay@dk

(* Repressor *)
let rep(r:chan()) =
  ?r; rep(r)

(* Negp gate *)
let negp(a:chan(), (cst:float, inh:float), p:proc()) =
  do ?a; delay@inh; negp(a,(cst,inh),p)
  or delay@cst; (p() | negp(a,(cst,inh),p))

(* Wiring *)
new TetR @1.0: chan()            (* TetR protein *)
new LacI @1.0: chan()            (* LacI protein *)
new LambcI @1.0: chan()   (* LambcI protein *)
new GFP @1.0: chan()             (* GFP protein *)
new aTc @100.0: chan()           (* aTc inducer *)
new IPTG @100.0: chan()   (* IPTG inducer *)

(* Auxiliary definitions: negp products *)
let rtr_TetR_aTc() = rtr(TetR,aTc)
let rtr_LacI_IPTG() = rtr(LacI,IPTG)
```

```
let tr_LambcI() = tr(LambcI)
let tr_GFP() = tr(GFP)

(* D038 Circuit *)
val PT = (0.1, 0.25)   (* PT constitutive and inhibition rates *)
val PL2 = (0.1, 1.0)   (* PL2 constitutive and inhibition rates *)
val Plm = (0.1, 1.0)   (* Plm constitutive and inhibition rates *)

let tet() = negp(TetR, PT, rtr_TetR_aTc)
let lac() = negp(TetR, PT, rtr_LacI_IPTG)
let cI()  = negp(LacI, PL2, tr_LambcI)
let gfp() = negp(LambcI, Plm, tr_GFP)

run
( tet() | lac() | cI() | gfp()
(* | rep(aTc)     uncomment to test with aTc *)
(* | rep(IPTG)    uncomment to test with IPTG *)
)
(* D016 Circuit *)
val PT = (0.1, 0.01)   (* PT constitutive and inhibition rates *)
val PL1 = (0.1, 0.01)   (* PL1 constitutive and inhibition rates *)
val PL2 = (0.1, 0.01)   (* PL2 constitutive and inhibition rates *)
val Plm = (0.1, 0.01)   (* Plm constitutive and inhibition rates *)

let tet() = negp(TetR, PT, rtr_TetR_aTc)
let lac() = negp(LacI, PL1, rtr_LacI_IPTG)
let cI()  = negp(LacI, PL2, tr_LambcI)
let gfp() = negp(LambcI, Plm, tr_GFP)

run
( tet() | lac() | cI() | gfp()
(* | rep(aTc)     uncomment to test with aTc *)
(* | rep(IPTG)    uncomment to test with IPTG *))
```

**Complexation**

Complexation can be modeled in stochastic process calculi by using a technique originally developed by Aviv Regev and Ehud Shapiro [6][7]. This technique provides a simple illustration of a major feature of process calculi that we have not emphasized in the main text: the dynamic creation of fresh communication channels. A fresh (unique) channel can be dynamically created, operationally, by incrementing a global counter, or by picking a random number. Process calculi abstract from these operational details by a formalized notion of what it means for a channel to be *fresh*. The operator *new* $c_r$; *P* creates a fresh channel named *c* with rate *r* for use in *P* (distinct from any other channel that might also be named *c*).

We want to model two proteins *P* and *Q* that combine into a complex *P:Q* at some rate *r*, and break apart again at some rate *s*. Let *cx* denote the complexation interaction of the two proteins: this is modeled as a single "public" channel *cx* of rate *r*, where multiple copies of *P* and *Q* can interact to come together and form complexes. Let *dx* denote the decomplexation interaction of two bound proteins: this is modeled as a separate channel *dx* of rate *s* for each complex. Such a fresh channel is established separately for each complex at the time of complexation, for the purpose of subsequently breaking up.

$P = new\ dx_s\ !cx(dx);\ !dx;\ P$
$Q = ?cx(x);\ ?x;\ Q$ where $x$ is an input variable

If we consider just one copy of $P$ and one of $Q$, for simplicity, the initial system $P|Q$ consisting of two separate proteins can evolve by $P$ creating a fresh channel $dx$ and outputting this $dx$ over the public channel $cx$, where it can be input by $Q$ and bound to its input variable $x$. At this point the system has evolved into the configuration $new\ dx_s\ (!dx;\ P) \mid (?dx;\ Q)$, where $dx$ is unknown to any other actual or potential process in the system. This state represents the complex of the original $P$ and $Q$. Next, an interaction can happen over this particular $dx$ channel among the only two processes that share it: this is the decomplexation event resulting in the initial state $P|Q$.

$$P|Q \ \rightarrow_r\ new\ dx_s\ (!dx;\ P) \mid (?dx;\ Q) \ \rightarrow_s\ P|Q \qquad \text{where } dx \text{ is fresh}$$

Many variations on this theme are possible, including modeling the binding, unbinding, and cooperative binding of transcription factors.

### Neg Gate Dynamic Response Profile

We test the dynamic response profile of the *neg* gate of Figure 3. To observe some of its behavior under operating conditions, we provide an input consisting of a signal raising linearly from 0 to 100, and then falling linearly from 100 to 0. That means 100 copies of input molecules, where each molecule is injected at a certain time and can interact or decay a certain number of times (thus shaping the input curve).

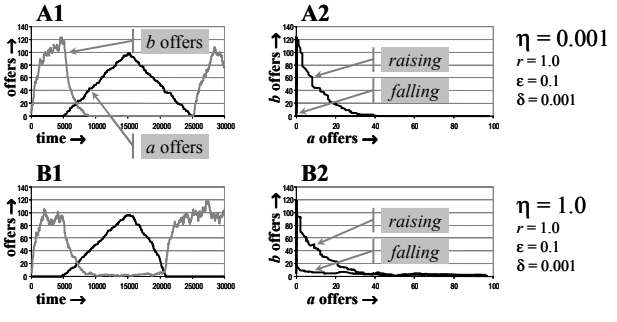



**Fig. 18.** Neg Gate Response Profile

Initially, in absence of any input, the output of the *neg* gate quickly raises to about 100. As the input signal ramps up, the output signal decays, and as the signal ramps down the output rises again, but with an asymmetric profile. (Figure 18 (A1,B1): the ramping down of the input signal in B1 appears abbreviated because the signal is consumed at a higher rate by the gate.) Plotting input vs output for the same data (Figure 18 (A2,B2)) we can see a roughly hyperbolic response with two distinct curves corresponding to raising and falling inputs. We show the plots for a highly sensitive ("Boolean") gate with $\eta$=0.001 (Figure 18 (A1,A2)) and a less sensitive gate with $\eta$=1.0 (Figure 18 (B1,B2)); these parameters cover the range used in simulations

in the main text. As in the main text, what is actually plotted is the number of (output) communication *offers* on the channels.

These response profiles illustrate the fact that, e.g., in the repressilator, each signal dynamically shapes the next signal and is shaped by the intake of the next gate.

The following is the complete code used to obtain the graphs, for the SPiM simulator (v0.04).

```
(* Simulation time, samples, and plotting  *)
directive sample 30000.0 1000
directive plot !a as "a"; !b as "b"

(* Parameters *)
val dk  = 0.001   (* Output protein decay rate *)
val inh = 0.001   (* Inhibition rate, or 1.0 *)
val cst = 0.1     (* Constitutive rate *)
val bnd = 1.0     (* Protein binding rate *)

(* Transcription factor *)
let tr(p: chan()) = do !p;tr(p) or delay@dk

(* Neg gate *)
let neg(a:chan(), b:chan()) =
  do ?a; delay@inh; neg(a,b)
  or delay@cst; (tr(b) | neg(a,b))

(* Probe signal: linearly raising and falling *)
val pbdk = 0.1    (* Probe signal decay rate *)
let probe1(p:chan(),n:int) =
  if n=0 then ()
  else (do !p;probe1(p,n-1) or delay@pbdk; probe1(p,n-1))
let dprobe1(p:chan(),d:int,n:int) =
  if d=0 then probe1(p,2*10*n)
  else delay@pbdk;dprobe1(p,d-1,n)
let probe(p:chan(),m:int) =
  if m=0 then ()
  else (dprobe1(p,500+(10*m),100-m) | probe(p,m-1))

(* Probing *)
new a@bnd:chan()  new b@bnd:chan()
run (neg(a,b) | probe(a,100))
```

# A Weighted Profile Based Method for Protein-RNA Interacting Residue Prediction

Euna Jeong and Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan
{eajeong, miyano}@ims.u-tokyo.ac.jp

**Abstract.** The prediction of putative RNA-interacting residues in proteins is an important problem in a field of molecular recognition. We suggest a weighted profile based method for predicting RNA-interacting residues, which utilizes the trained neural network. Most neural networks have a learning rule which allows the network to adjust its connection weights in order to correctly classify the training data. We focus on the network weights that are dependent on the training data set and give evidence of which inputs were more influential in the network. A large set of the network weights trained on sequence profiles is analyzed and qualified. We explore the feasibility of utilizing the qualified information to improve the prediction performance for protein-RNA interaction. Our proposed method shows a considerable improvement, which has been applied to the profiles of the PSI-BLAST alignment. Results for predictions using alternative representations of profile are included for comparison.

## 1 Introduction

Structural studies of protein-RNA complexes have been investigated at diverse aspects, such as analyses of specific RNA recognition mode in proteins [7], the binding properties in protein-RNA interface [15], the chemical principles governing both specific and non-sequence specific binding [1], the atomic and amino acid level properties with secondary structural effect in hydrogen bonding [17], and the energetic features in protein-RNA recognition [4]. The interaction patterns discovered from the analyses give us useful information to understand specificity and affinity in protein-RNA interaction. However, very few studies have been addressed so far to the important problem of predicting RNA-interacting sites in proteins. One of the main reasons is that there was a small number of interactions available for training a learning system. An appropriate approach to the prediction of interactions relies on enough training data from the statistical examination observed in high-resolution crystal structures.

There is a related approach which is based on support vector machines to recognize whether a given protein is RNA-interacting or not [9]. This work is different from ours in the target of prediction as identifying RNA-binding proteins, not RNA-interacting residues in proteins. In addition, it may be argued that their result has biased by redundant homologous structures, because duplicated

C. Priami et al. (Eds.): Trans. on Comput. Syst. Biol. IV, LNBI 3939, pp. 123–139, 2006.

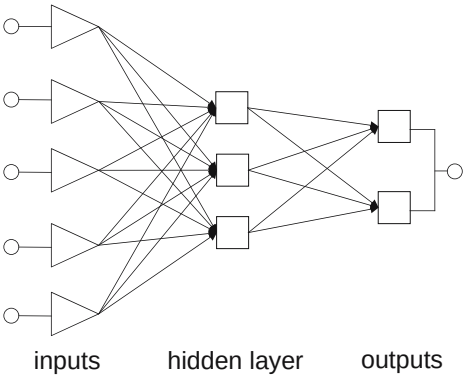



**Fig. 1.** An example of a simple feedforward network

structures are only eliminated from the dataset. Recently we proposed a neural network based method for predicting RNA-interacting residues given a protein chain that is already known to form a complex with an unknown RNA [13]. The system used the protein sequence information and the corresponding secondary structure information as input of the neural network. The approach was motivated to examine the correlation between the segments of the consecutive residues with similar secondary structure states and their roles in RNA recognition. The comparison of performance will be described in Section 3.2.

In this study, we suggest a weighted profile based approach, which utilizes the trained neural network. Profile is a substitution matrix which can be deduced from given aligned sequences. Many studies have used the neural networks trained on sequence profiles for classifying of protein secondary structure [6, 14, 18, 23], of $\beta$-turn types in proteins [16], of solvent accessible surface [24], and of protein-protein interaction [31]. We discuss a method to analyze and qualify a large set of the network weights which are trained on sequence profiles, and explore the feasibility of utilizing the qualified information to improve the prediction performance for protein-RNA interaction.

A typical feedforward network has units (or neurons) arranged in a distinct layered topology as shown in Figure 1. Units are connected to one another and connections correspond to the edges of the directed graph. There is a real number associated with each connection, which is called the weight of the connection. An important component of most neural networks is a learning rule which allows the network to adjust its connection weights in order to correctly classify the training data. We focus on the network weights that are dependent on the training data set and give evidence of which inputs were more influential in the network.

The saliency factor is defined to represent the amount of information embedded in the network weights. Given a network trained on sequence profiles, we assume that the magnitude of a weight reflects the significance of the presence of a specific amino acid at a specific position. For example, the weight whose magnitude is small will have a minor effect on the performance of the network, while the large magnitude of the weight will be a great influence on it. In order

to view the usefulness of our proposed method, a multi-layered neural network is trained on profiles available from protein families. The saliency factor is derived from the trained network first. The sequence profiles are weighted by the saliency factor and then used as input of the neural network. Results of alternative representations of profile are also examined for comparison. Experimental results demonstrate that our proposed method has benefits.

## 2 Materials and Methods

The prediction process of RNA-interacting residues is divided into five stages: (1) generation of multiple sequence profiles, (2) training of the neural networks based on the profiles, (3) calculation of a saliency factor, (4) generation of weighted profiles, and lastly (5) prediction of RNA-interacting residues using the weighted profiles, as shown in Figure 2.

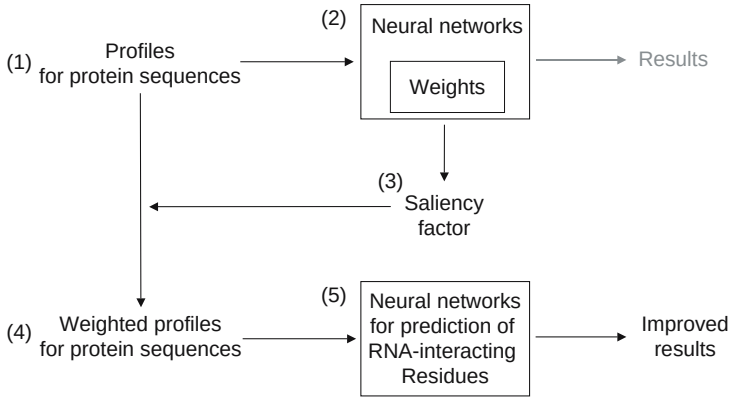



**Fig. 2.** Diagram of the prediction process

### 2.1 Dataset Generation

We examined 87 non-homologous protein chains from the same dataset of our earlier work [13]. Protein sequences used for the network training were obtained from PDB [3], which were resolved by X-ray crystallography with better than 3.0 Å resolution. Redundant protein structures with sequence identity over 70% are removed after homologous sequence search by the PSI-BLAST program [2]. For this study, protein sequences were screened to remove proteins which had interacting residues less than 4, and those which did not generate valid alignment profiles. The selected 87 protein chains are assumed to form interactions with one RNA chain which contains at least 4 nucleotides in length.

We consider protein-RNA interactions which include hydrogen bonding, stacking, electrostatic, hydrophobic and van der Waals interactions. Distances within 7.0 Å was used to classify diverse interactions including electrostatic interactions, since those distances can be important to protein-nucleic acid interactions [1, 27].

We chose 6.0 Å as a cutoff to define a wide range of protein-RNA interactions in consideration of experimental noise. A residue is considered to be in RNA-interacting if the closest distance between atoms of the protein and the partner RNA is within the cutoff. If a protein chain is common to two or more protein-RNA sequence pairs, the protein sequence which is resolved with a higher resolution or which has the largest number of RNA-interacting residues is selected.

### 2.2 Profile

The group of related sequences is aligned together to create a multiple sequence alignment, which reveals if there is an evolutionary relationship between the sequences. Given an aligned set of sequences, each position in the alignment can be represented quantitatively as a vector of 20 scores derived from position-specific symbol comparison. This is called a profile.

Generally the score for a specific amino acid at a specific position is the real counts plus pseudocounts all divided by the total number of possible counts: $(q_a+p_a)/(Q+P)$ where $q_a$ and $p_a$ are the real counts and pseudocounts, respectively, of amino acid $a$ and $Q$ and $P$ are the total number of real counts and pseudocounts, respectively, in the position. Pseudocounts are introduced to improve the estimates of the amino acid frequencies because the alignment is a sample drawn from a much larger distribution and most counts are zero [12]. Therefore the profile is influenced by the number of sequences in the alignment and their diversity. In addition, the score of a specific amino acid at a specific position is alternatively represented according to the methods for estimating the expected amino acids. To compare the difference between the algorithms used to make optimal alignments, as well as scoring methods, we use two types of alignment – local and global alignment, and three scoring methods – no pseudocounts, substitution matrices, and Dirichlet mixtures [28].

The PSI-BLAST [2] and CLUSTALW [29] programs are considered to generate local and global alignment, respectively. PSI-BLAST builds alignment that includes sequences with more remote similarities by using iterative search. CLUSTALW uses a progressive alignment method starting with the most related sequences and then sequentially adding less-related sequences to the initial alignment. The PSI-BLAST search of a query sequence is carried out against NR database to collect the proteins in a family, which was used for three iterations at e-value 0.0001. The result alignment of PSI-BLAST is prepared for input of CLUSTALW, which was used with default values. It is intended to obtain two different alignments for the same set of related sequences and hence to compare the effect of two alignment programs.

From two kinds of sequence alignment, alternative representations of profiles are considered by using three different scoring methods: (1) a frequency profile based on a simple formula with no pseudocounts; (2) a PSI-BLAST position specific scoring matrix (PSSM) based on amino acid substitution matrices; and (3) a profile HMM [19] based on Dirichlet mixtures using the previous variations observed in the BLOCKS database [11]. The frequency profile is calculated as a ratio of the occurrences for a particular amino acid and the total number

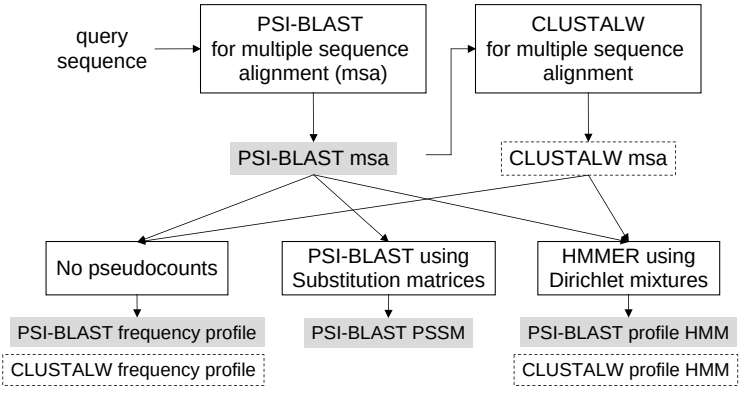



**Fig. 3.** The generation of profiles using two types of alignment and three different scoring methods. The lined box shows the used tools (or methods) and the arrow means the flow of data. The grayed and dotted box show the result generated from PSI-BLAST and CLUSTALW alignment, respectively.

of sequences at a particular position, i.e., $p_a = P = 0$, similar to the method applied in [6]. The PSI-BLAST PSSM is an intermediate matrix generated as part of the search process and we directly use it as a profile. The third one is the profile HMM represented by a hidden Markov model (HMM). It is created by the HMMER [8] package from a set of aligned sequences. The generation of profile is illustrated in Figure 3.

### 2.3 Neural Network Modeling

We use a set of feed-forward neural networks trained by back-propagation. The publicly available simulation package SNNS 4.2 [30] is used to implement the neural networks.

A window of length 15 residues is used to move along the protein sequence, and the network is trained to predict whether the central residue of the window interacts with RNA or not. The 20 amino acids are encoded in a position-specific scoring matrix. The matrix contains $15 \times 21$ elements, and each element represents a particular residue substitution at a particular position in the sequence. An additional input is required per residue for windows overlapped either end of the sequence.

Normalization of the input data is needed for a neural network to operate properly, which reduces computational time. This procedure is also required to scale out these input sets in order to have cross-profile comparative analysis. For the frequency profile, the score is normalized by multiplying by ten and rounded for its simplicity, and for PSI-BLAST PSSM and profile HMM the range is between 0 and 1 by using the standard logistic function.

A fully connected, two-layer feed-forward network with 315 input ($15 \times 21$), 5 hidden, 1 output neuron is used. It is selected as an optimal network after various network topologies were explored with the combination of different sizes

of window ranging from 7 to 45 and different numbers of hidden unit tuning from 1 to 9. The best results are decided by the mean squared error (MSE) as a criterion. The target of the output neuron is 1 if the central residue belongs to an RNA-interacting site and otherwise, 0 is used.

Each of input patterns is presented to the networks in randomized order and at least chosen once during the training process. Evaluating a prediction method is done by a 10-fold cross-validation where the dataset is divided into 10 subsets. All simulations use the learning factor of 0.1 and momentum term of 0.001.

### 2.4 Saliency Factor

We define the saliency factor as a matrix to represent the importance of the presence of specific residues at specific positions. The larger the magnitude of a weight is, the higher its level of participation in the solution is. On the other hand, if the value of a weight is near zero, the connected input unit may be not necessary to the solution. In this study, since the experiments were designed to investigate the significance of input units, we focus the weights connecting the input layer with the hidden layer.

We use the notation summarized in Table 1. Matrices are indicated with calligraphic letters, whose size is all $M \times N$ where $M$ is the window size and $N$ is the number of distinct amino acids.

The input layer reads input pattern $\mathcal{A}$ that is coded as a numeric string consisting of 15 blocks of 20 elements (an additional input is omitted for analysis)

**Table 1.** Notation

| | |
|---|---|
| $M$ | the window size |
| $N$ | the number of distinct residue symbols |
| $H$ | the number of hidden units |
| $i$ | the variable for $1 \leqslant i \leqslant M$ |
| $j$ | the variable for $1 \leqslant j \leqslant N$ |
| $k$ | the variable for $1 \leqslant k \leqslant H$ |
| $\mathcal{A}$ | the input pattern whose size is $M \times N$ |
| $\mathcal{F}$ | the saliency factor whose size is $M \times N$ |
| $\mathcal{P}$ | the new input pattern whose size is $M \times N$ |
| $a_{ij}$ | the element of $\mathcal{A}$ |
| $f_{ij}$ | the element of $\mathcal{F}$ |
| $p_{ij}$ | the element of $\mathcal{P}$ |
| $h_k$ | the hidden unit |
| $w_{ijk}$ | the weight from $a_{ij}$ to $h_k$ |
| $x_{ij}$ | the normalized weight for input unit $a_{ij}$ |
| $R_i$ | the weight conservation at window position $i$ |
| $X_i$ | the sum of $N$ normalized weights at window position $i$ |
| $E_i$ | the entropy of the observed weight distribution at window position $i$ |

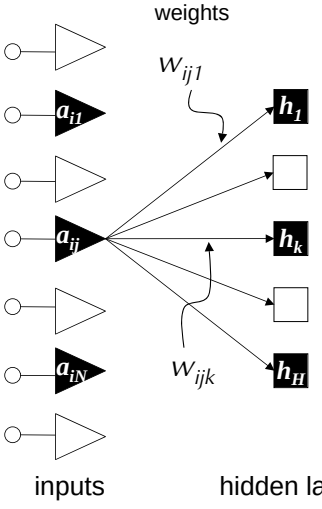



**Fig. 4.** Illustration of a part of the neural network. Given a window of length $M$ residues, a residue at window position $i$ is encoded with $N$ amino acid substitution scores ($a_{i1}$, $\cdots$, $a_{iN}$) for $1 \leqslant i \leqslant M$ and $1 \leqslant j \leqslant N$. Each input unit $a_{ij}$ has $H$ weights ($w_{ij1}$, $\cdots$, $w_{ijH}$) connecting with hidden unit $h_k$, for $1 \leqslant k \leqslant H$. An open polygon represents an arbitrary number of units and for convenience other connections are omitted.

and is fully connected to the 5 hidden units. Figure 4 shows a small part of the network focused on the $i$th position in the sliding window, which is encoded with $N$ amino acid substitution scores, and the $H$ weights of the connection between input node $a_{ij}$ and hidden node $h_k$ for $1 \leqslant i \leqslant M$, $1 \leqslant j \leqslant N$, and $1 \leqslant k \leqslant H$. The number of weights analyzed is totally $M \times N \times H$.

We first calculate the normalized weight $x_{ij}$ for each input unit $a_{ij}$ by dividing the summation of the absolute values of weights by the total number of hidden units.

$$x_{ij} = \frac{\sum_{k=1}^{H} |w_{ijk}|}{H}. \tag{1}$$

Since the neural networks are trained using a 10-fold cross-validation procedure, Equation 1 is averaged over 10 sets.

Saliency factor $\mathcal{F}$ is derived from qualification of weight conservation similar to the sequence conservation defined by Schneider and Stephens [26]. The weight conservation is the amount of weight information present at each position $i$ in the given window. It is defined as the difference between the maximum entropy and the entropy of the observed weight distribution:

$$R_i = \log X_i - E_i \tag{2}$$

in which

$$X_i = \sum_{j=1}^{N} x_{ij} \text{ and } E_i = -\sum_{j=1}^{N} (\frac{x_{ij}}{X_i}) \log(\frac{x_{ij}}{X_i}). \tag{3}$$

Element $f_{ij}$ of $\mathcal{F}$ is calculated by multiplying the observed weight distribution of residue $j$ by the weight conservation at window position $i$, derived from Equations 1 and 2,

$$f_{ij} = \frac{x_{ij}}{X_i} R_i \tag{4}$$

Given input pattern $\mathcal{A}$ and saliency factor $\mathcal{F}$, new input pattern $\mathcal{P}$ is constructed by multiplying elements of $\mathcal{A}$ and $\mathcal{F}$ at the same position. New input unit $p_{ij}$ of $\mathcal{P}$ is defined as

$$p_{ij} = a_{ij} \ f_{ij}. \tag{5}$$

### 2.5 Measures of Performance

The performance of a particular prediction is assessed by total accuracy, accuracy, coverage (sensitivity), specificity and Mattews correlation coefficient, as follows:

- Total accuracy as the percentage of all correct predictions,

$$\frac{tp + tn}{tp + tn + fp + fn} \times 100,$$

- Accuracy as the ratio of correctly predicted interaction sites and predicted interaction sites,

$$\frac{tp}{tp + fp} \times 100,$$

- Coverage, also called sensitivity, as the ratio of correctly predicted interaction sites and observed interaction sites, defined by

$$\frac{tp}{tp + fn} \times 100,$$

- Specificity as the ratio of correctly predicted not-interaction sites and observed not-interaction sites,

$$\frac{tn}{tn + fp} \times 100,$$

- Matthews correlation coefficient (MCC) as a more complicated measure indicating the magnitude of correlation between actual and predicted values, using

$$\frac{(tp)(tn) - (fp)(fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}}$$

where $tp$ is for the number of correctly predicted interacting residues, $tn$ for that of correctly predicted not-interacting residues, $fp$ for that of actually not-interacting residues predicted as interacting one, and $fn$ for that of actually

interacting residues predicted as not-interacting one. Sensitivity and specificity are considered for the receiver operating characteristic (ROC) as a threshold independent measure. The MCC is used as a major measure because it takes account of both under- and over-predictions. The MCC value is ranged between 1.0 and -1.0, which corresponds to a perfect and a completely wrong prediction, respectively.

## 3 Experimental Results

We first report the prediction results from the neural networks based on the profiles (i.e., from the second stage in Figure 2) and then the results from the neural networks based on weighted profiles (i.e., from the last stage in Figure 2). In addition, analyses of the saliency factor are given lastly.

### 3.1 Prediction Based on Profiles

We generated three profiles from the PSI-BLAST alignment and two profiles from the CLUSTALW alignment, respectively. The prediction results using different alignment methods combined with different scoring methods are summarized in Table 2.

For the same sequences, the predictions from the PSI-BLAST alignment appear to be better than those from the CLUSTALW alignment. The MCC of PSI-BLAST PSSM is 0.39. The PSI-BLAST frequency profile and the PSI-BLAST profile HMM show the same MCC as 0.35. Although the MCC values are same, in detail, the accuracy and coverage of the profile HMM are 4% less and 4.9% better than the frequency profile, respectively.

In the case of CLUSTALW, the MCC value of the frequency profile is 0.33. The predictions of the frequency profile show similar performance in two alignment programs. The MCC value from the PSI-BLAST alignment is found to be 0.02 better than that from the CLUSTALW alignment. However, the profile HMM from the CLUSTALW alignment shows MCC of 0.17, around half of the CLUSTALW frequency profile, which is the least performance.

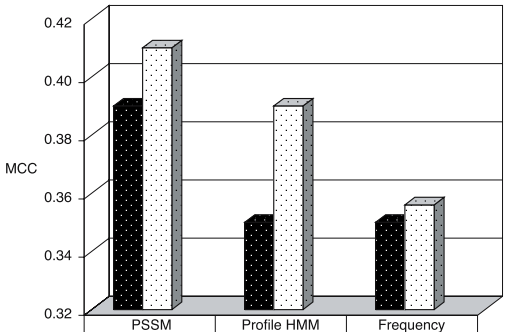
As compared with other profiles, the PSI-BLAST PSSM obtains the best result as total accuracy of 80.2%, accuracy of 58.9%, coverage of 43.4% and MCC of 0.39.

**Table 2.** Comparison of prediction performance between different profiles

| Alignment | Scoring matrix | Total accuracy | Accuracy | Coverage | MCC |
|---|---|---|---|---|---|
| PSI-BLAST | frequency profile | 79.8 | 59.1 | 36.9 | 0.35 |
| | PSSM | 80.2 | 58.9 | 43.4 | 0.39 |
| | profile HMM | 79.0 | 55.1 | 41.8 | 0.35 |
| CLUSTALW | frequency profile | 79.0 | 56.1 | 35.3 | 0.33 |
| | profile HMM | 75.6 | 42.6 | 21.1 | 0.17 |

### 3.2 Effect of Profiles Weighted by the Saliency Factor

We focus on the trained networks based on the three PSI-BLAST profiles to verify the effect of the saliency factor, since the PSI-BLAST profiles show better performance than the CLUSTALW profiles.

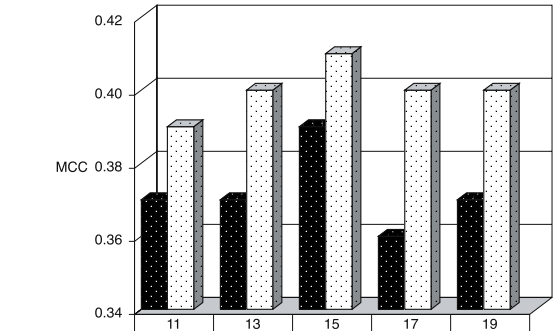



| | PSSM | Profile HMM | Frequency |
|---|---|---|---|
| PSI-BLAST profile | 0.39 | 0.35 | 0.35 |
| Weighted profile | 0.41 | 0.39 | 0.36 |
| Improvement | 0.02 | 0.04 | 0.01 |

**Fig. 5.** Comparison between predictions of the profiles generated from the PSI-BLAST alignment and the weighted profiles

The saliency factor is derived from the trained network based on each profile. The profile is weighted by its saliency factor and the weighted profile is used to train the neural network with the same topology as described in Section 2.3. The results are shown in Figure 5. The black and white bars show the MCC values achieved by the profiles and the weighted profiles, respectively. The corresponding values are shown in the below of the bars with the improvement obtained after using the weighted profiles. As using the weighted profiles, the performance is improved as a whole. The MCC values rose from 0.39 to 0.41 with the PSSM, from 0.35 to 0.39 with the profile HMM, and from 0.35 to 0.36 with the frequency profile, respectively. The PSSM weighted by its saliency factor also shows the best result among the three profiles. It is interesting that the most improvement is achieved by the weighted profile HMM, while the weighted frequency profile shows the least improvement in this case. We have observed that the amount of information contained in the profile HMM is much more than that in the frequency profile, in a sense that a weighted profile is a profile weighted by its carrying information.

We also examined the relationship between the weighted profiles and the neural networks with different topology. The neural networks are trained on the weighted PSI-BLAST PSSM with different window size ranging from 11 to 19. Figure 6 summarizes the results. The table format is the same as Figure 5 and each column in the table represents different window size, rather than different profiles. After using our method, the improvement is also shown in all ranges of window size. We have obtained final MCC values of 0.39, 0.40, 0.41, 0.40 and



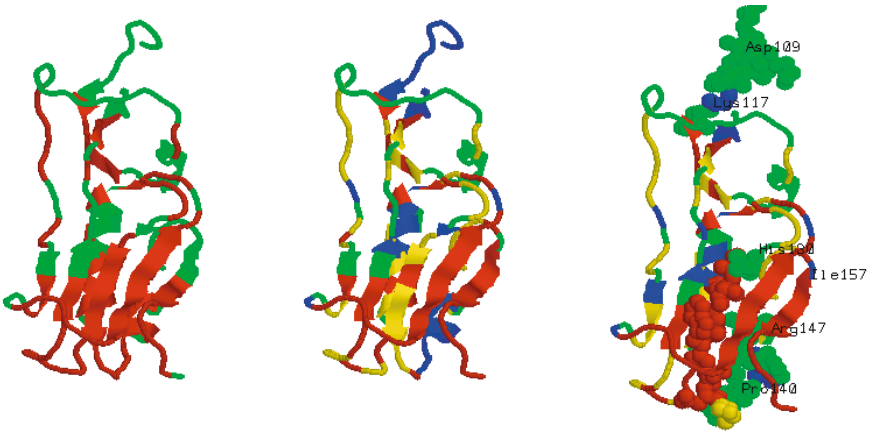

| | 11 | 13 | 15 | 17 | 19 |
|---|---|---|---|---|---|
| ■ PSI-BLAST PSSM | 0.37 | 0.37 | 0.39 | 0.36 | 0.37 |
| □ Weighted PSSM | 0.39 | 0.40 | 0.41 | 0.40 | 0.40 |
| Improvement | 0.02 | 0.03 | 0.02 | 0.04 | 0.03 |

**Fig. 6.** Comparison between predictions of the PSI-BLAST PSSM and the weighted PSSM along with different window sizes

0.40, in the increasing order of window size, respectively. The MCC values are all raised to $\geqslant 0.39$ which was the best MCC obtained from PSI-BLAST PSSM prediction at window size 15 before weighting. We have two observations that (1) the network architecture with window size 15 is also optimal when using the weighted PSSMs, and (2) the performance between the weighted PSSMs becomes similar, since the difference between MCC values is slightly reduced after weighting. The saliency factor is basically dependent on the input data. Even if different network architectures are used, the network input carries similar amount of information because of using the same profile.

As compared with single sequences based predictions in a previous work [13], the discriminating capability of the predictor is significantly improved by using the weighted profile, in particular, the weighted PSI-BLAST PSSM. Using single sequences information combined with secondary structure information obtained MCC of 0.29 at window length 41. The prediction performance of the weighted PSSM leads to a significant increase of 0.12 in the MCC value. As considered the window size, the overall performance is much better than 0.12.

Figure 7 depicts the difference between two predictions, by using the PSSM and the weighted PSSM for ribosomal protein 1jj2:H. In the details of predictions, the differences obtained after applying our method are represented by C for better predictions, I for worse predictions, and space for no changes. This result shows that our proposed method can distinguish between interacting and not-interacting sites with more accuracy. It is observed that the better changes, denoted as C, occurred consecutively, while the worse changes happened sparsely. The Cs are concentrated in the ranges of residue number Asp109 - Lys117, His130 - Pro140, and Arg147 - Ile157. These ranges are emphasized by using ball-and-stick format in (c). It is possible that the strength of the network to capture the correlation between consecutive residues is amplified by using weighted profiles.




(a) observed intraction (b) predicted by the PSSM (c) predicted by the weighted PSSM

```
          81        91     109         121       131       141       151       161
Sequence  YKFRIRKFPFHVIRENKDGMRAPFGKPVGTAARVHGANHIFIAWVNPDPNVEEAWRRAKMKVTPTINIDSSPAGNA
     (a)  00011111010001111111111111000001000011000000000000001001101111111110000000001
     (b)  00010111111000000000000000010101101111011011111110000000100000000011111111101
     (c)  10010101111000010111111101000101101111100000000000000000111111110111101111100
Difference I     I        C CCCCCCC C I          CCC CCCCCCC      C ICCCCC CC  C      I
```

(d) predictions in detail

**Fig. 7.** Predictions for ribosomal protein sequence 1jj2:H drawn by application RasMol [25]: (a) the observed interactions; (b) the predicted by the PSSM; (c) the predicted by the weighted PSSM; (d) a portion of predictions in detail. Colors for proteins: green for interaction sites; red for not-interaction sites; yellow for over-predictions; blue for under-predictions. In (d), the observed and predicted results are represented by 1 for interactions and 0 for not-interactions. The differences between two predictions (b) and (c) are denoted by C for changed correctly by our proposed method, I for changed incorrectly by ours and space for the same predictions. Ball-and-stick format in (c) is used to emphasize some parts of difference between two predictions.

### 3.3 Information Contents in the Saliency Factor

The performance improvement of our proposed method depends on the information contents in the saliency factor that reflects the correlation between the window position and types of amino acid. The saliency factor calculated for the networks trained on the PSI-BLAST PSSM, is displayed in Figure 8. The $x$-axis shows each position in the window and the $y$-axis represents the amount of weight conservation of each amino acid at each position, which is the value of $f_{ij}$ in Equation 4.

The curves give evidence of what composition of amino acid types is contributed to the classification of the neural networks. The decreasing order of weight conservation at window position 8 is Glu, Ala, Asp, Arg, Leu, Val, Gln, Ser, Thr, Phe, Lys, Ile, Asn, His, Tyr, Pro, Met, Cys, Gly, and Trp. Basic and acidic amino acids excluding His show all high levels of distributions in the center of the window. Glu content is high in the right-hand side, and Asp shows

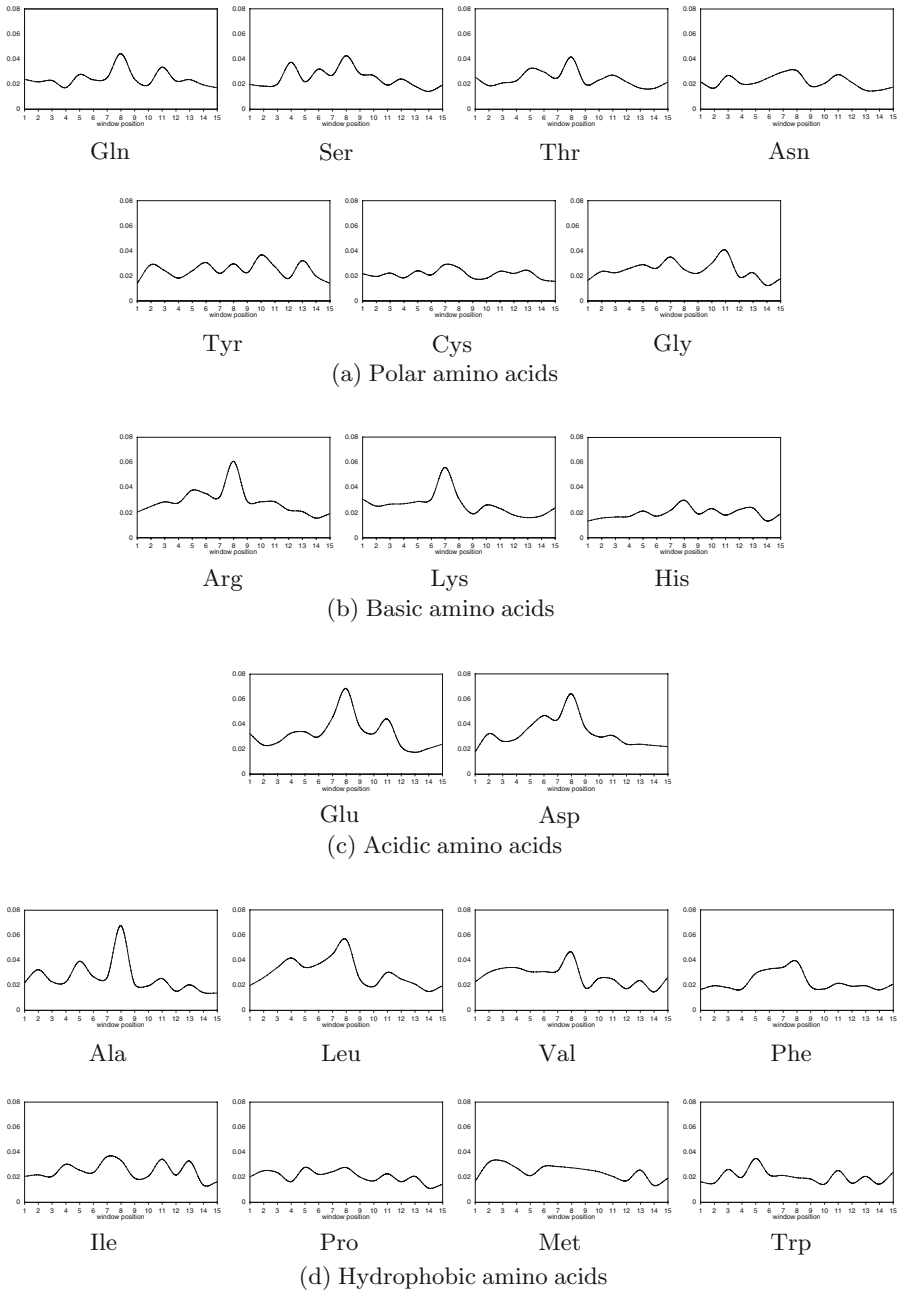



**Fig. 8.** Weight conservation for each of 20 amino acids according to the chemical species they represent [5]; (a) polar, (b) basic, (c) acidic, and (d) hydrophobic amino acids

high distribution in the left-hand side. Lys content in the center position is not too high, but it is high in the left-hand side of the center. High levels of content are also observed for hydrophobic residues Ala, Leu and Val, which are known to less preferred types in protein-RNA interactions. These results are in agreement with structural studies of protein-RNA interactions. Amino acid types Arg, Lys, Ser, Thr, Asp and Glu prefer to be in hydrogen bonding [1, 15, 17], and Phe and Ser are frequently located in van der Waals interacting and stacking interacting [1, 15]. There is also a conflicting situation. For example, Asn content does not contribute too much to classify the central amino acid as RNA-interacting residues, though typically thought of one of the most preferred amino acid types in hydrogen bonding.

### 3.4 ROC Curves

The receiver operating characteristic (ROC) is used to assess performance of networks trained on different profiles. Figure 9 shows the ROC curve for each of profiles. How well a test performs is represented by the area under the ROC curve. The closer the area is to 1.0, the better the test is. The corresponding areas under the curves are 0.77 by the weighted PSSM, 0.75 by the PSSM, 0.71 for the frequency profile, 0.69 by the profile HMM of PSI-BLAST, respectively. From the CLUSTALW alignments, the areas are 0.67 by the frequency profile and 0.57 by the profile HMM. We can see that the prediction using the weighted PSSM is the best among other predictions and the predictions from the PSI-BLAST alignments are more accurate for the same cross-validated test than those from the CLUSTALW alignments.

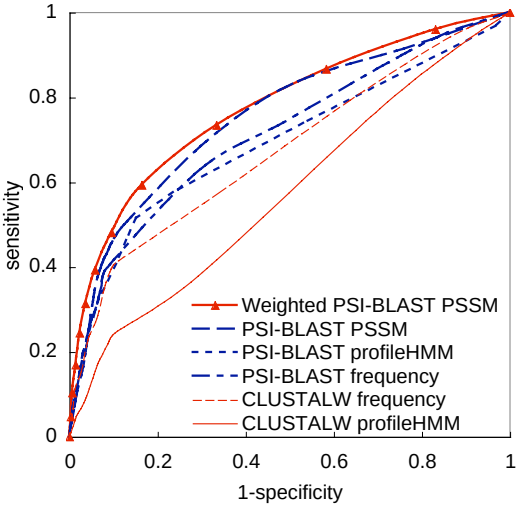



**Fig. 9.** ROC curves for the neural networks trained on different profiles at window size 15

## 4 Discussion and Conclusions

We have presented a weighted profile based approach for RNA-interacting residue prediction. This approach suggests a novel method to utilize information embedded in the weights of the neural networks. The experimental results show that using profiles weighted by its information, i.e., the saliency factor, improves prediction performance. In particular, when applied to the PSI-BLAST PSSM, the MCC value is raised to 0.41 which is the best achieved so far. It demonstrates that our proposed method increases the network's ability to find more relations.

The profiles generated from the PSI-BLAST alignment shows generally better performance than the profiles from the CLUSTALW alignment. It corroborates the result that the use of reliable local alignments produces a definite improvement in the accuracy of resulting secondary structure predictions [6, 14].

We also observed that the profile HMM from the CLUSTALW alignment shows the least performance. It has been reported that the profile HMM from re-scoring the CLUSTALW alignment could improve total accuracy as compared with the frequency profile in protein secondary structure prediction [6]. Our result cannot be directly compared with their result because of different domains. From the viewpoint of the total accuracy, however, the prediction achieved by the profile HMM of CLUSTALW is 3.4 - 4.6% lower than the results from other profiles. It is not clear which factors decrease the performance of the profile HMM. One limitation of HMMs concerns in which a range of training set sizes is crucial for successful learning. In order to produce a useful model for the sequences, HMMs require a quite large number of training examples, 20 or more [21]. In our dataset, around 9% of 87 protein chains has 20 or less sequences in the alignment. In addition, CLUSTALW performs pairwise global alignments of all of the sequences. The main problem of CLUSTALW is that the result alignment is heavily dependent on the initial pairwise sequence alignments. Any few errors in the initial alignments may be propagated to the result alignment [22]. It seems that a combined effect of the CLUSTALW alignments and HMMs causes the worst result.

It is known that there is a strong correlation between RNA-interacting residues and their compositions. However, it is not clear how the primary sequence of protein recognizes the RNA with specificity. As noted in Allers and Shamoo [1], the reason may be found from our poor understanding of protein folding, rather than RNA recognition. There are already known RNA-binding domains and motifs to mediate RNA recognition by proteins, such as RNA-recognition motif (RRM), double-stranded RNA-binding motif (dsRM), K-homology (KH) domain and S1 RNA-binding domain. The number of protein sequences containing these domains is very few, only around 12% of RNA-binding proteins [9]. It means that more efforts are needed to identify the molecular determinants of RNA recognition by proteins. The analysis of the weight distributions from our approach can be contributed to understand spatial and chemical features about consecutive residues in RNA-interacting sites.

In future work, design techniques for minimizing the size and structure of a neural network will be useful to make the saliency factor more optimal by

removing unimportant weights from the network [10, 20]. It also allows for finding generic principles underlying RNA-interacting sites more efficiently.

## Acknowledgments

We thank Keun-Joon Park and Michiel de Hoon for valuable discussions. Computer time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo for the use of supercomputer facilities.

## References

1. J. Allers and Y. Shamoo. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.*, 311:75–86, 2001.
2. S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
3. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.
4. Y. Chen, T. Kortemme, T. Robertson, D. Baker, and G. Varani. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Res.*, 32:5147–5172, 2004.
5. G.E. Crooks, G. Hon, J.-M. Chandonia, and S.E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14:1188–1190, 2004.
6. J.A. Cuff and G.J. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, 40:502–511, 2000.
7. D.E. Draper. Themes in RNA-protein recognition. *J. Mol. Biol.*, 293:255–270, 1999.
8. S.R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.
9. L. Han, C. Cai, S. Lo, M. Chung, and Y. Chen. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, 10:355–368, 2004.
10. B. Hassibi and D.G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in Neural Information Processing Systems*, 5:164–172, 1993.
11. S. Henikoff and J.G. Henikoff. Protein family classification based on searching a database of blocks. *Genomics*, 19:97–107, 1994.
12. S. Henikoff and J.G. Henikoff. Using substitution probabilities to improve position-specific scoring matrices. *CABIOS*, 12:135–143, 1996.
13. E. Jeong, I. Chung, and S. Miyano. A neural network method for identification of RNA-interacting residues in protein. *Genome Informatics*, 15:105–116, 2004.
14. D.T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.
15. S. Jones, D.T.A. Daley, N.M. Luscombe, H.M. Berman, and J.M. Thornton. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, 29(4):943–954, 2001.

16. H. Kaur and G. Raghava. A neural network method for prediction of $\beta$-turn types in proteins using evolutionary information. *Bioinformatics*, 20:2751–2758, 2004.
17. H. Kim, E. Jeong, S.W. Lee, and K. Han. Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Lett.*, 552:231–239, 2003.
18. A. Kloczkowski, K.L. Ting, R.L. Jernigan, and G. Garnier. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins*, 49:154–166, 2000.
19. A. Krogh, M. Brown, I.S. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology. *J. Mol. Biol.*, 235:1501–1531, 1994.
20. Y. Le Cun, J.S. Denker, and S.A. Solla. Optimal brain damage. *Advances in Neural Information Processing Systems*, 2:598–605, 1990.
21. G. Mitchison and R. Durbin. Tree-based maximal likelihood substitution matrices and hidden Markov models. *J. Mol. Evol.*, 41:1139–1151, 1995.
22. W. Mount. *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001.
23. B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19:55–72, 1994.
24. B. Rost and C. Sander. Conservation and prediction of solvent accessibility in protein families. *Proteins*, 20:216–226, 1994.
25. R. Sayle and E.J. Milner-White. RasMol: Biomolecular graphics for all. *Trends Biochem. Sci.*, 20:374–376, 1995.
26. T.D. Schneider and R.M. Stephens. Sequence logos: a new way to display consensus sequence. *Nucleic Acids Res.*, 18:6097–6100, 1990.
27. K.A. Sharp, B. Honig, and S.C. Harvey. Electrical potential of transfer RNAs: codon-anticodon recognition. *Biochemistry*, 29:340–346, 1990.
28. K. Sjölander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci.*, 12:327–345, 1996.
29. J.D. Thompson, D.G. Higgins, and T.J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.
30. A. Zell and G. Mamier. Stuttgart neural network simulator version 4.2, 1997.
31. H.X. Zhou and Y. Shan. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins*, 44:336–343, 2001.

# Author Index