Symbiose

Biological Systems and Models, Bioinformatics and Sequences

Jacques Nicolas

https://www.irisa.fr/symbiose
Overall objectives

• A bioinformatics project:
  – From sequence or expression data to biological knowledge
  – Interest for BioInfo:
    Formal and predictive models in molecular biology
  – Associated bioinformatics platform.

• Large scale Bioinformatics.
  Focus on discrete, symbolic models
  – Optimization methods;
  – Interaction with biological labs, boosted via a bioinformatics platform.

• Research axes:
  ➢ High performance computing for bioinformatics
  ➢ Modelling sequence/structure relationships;
  ➢ Large scale modelling in Systems Biology.
Research

Codesign of algorithms and hardware architectures tailored to large scale bioinformatics: study of reconfigurable machines using FPGA or fast components such as Flash memories or GPUs.

Language theory and combinatorial optimization: efficient filters and model matching in large data banks; design and learn grammatical models on biological sequences; protein structure prediction.

Graph models of biological interaction networks & derivation of discrete or differential models explaining / predicting the behavior of the system. Diagnosis of large scale models.
Symbiose & GenOuest

- **Symbiose is the research team:**
  Head  J. Nicolas
  10 researchers (4 INRIA/CNRS, 6 Faculty)
  10 PhD, 3 post-docs, 3 engineers en CDD

- **Genouest is the computing center in bioinformatics**
  Head  O. Collin
  3 full-time engineers,
  5 engineers on contracts
Bioinformatics Platform
http://genouest.org

Welcome on GenOuest bioinformatics platform

Welcome on the OUEST-genopole® bioinformatics server

This platform aims at proposing innovative software in Bioinformatics. You will find here the latest software of the bioinformatics, public databases updated on a daily basis, and a range of links to seminars, training courses, platform news...

For a direct access to the tools, please use thumbnail "tools" Yes:

- Exclusive tools such as STAN or Wapam (under heading "Analysis") and Domain Organizer (under heading "Protocole")
- Specialized relational databases such as PUBMED (under heading "Analysis") and GenOuest Workbench (under heading "Protocole")

The Search toolbox (up right) provides an easy access to documents and tools available on this site.

If you wish to give access to your software (possibly a prototype without interface) or on the contrary, if you are looking for some functionalities and do not find any practical tool, please use the FAQ.

We propose you to subscribe to the OUEST-genopole® bioinformatics platform’s mailing-list : it will allow you to know news about software updates, events or trainings.

Platform events

Plateform news

25 October 2006 : "Nantes/Actualités/Transplantation"
Main collaborations

• International
  – Bulgaria, Univ. Sofia (N. Yanev)
  – Germany, Univ. Potsdam (T. Schaub)
  – China, Institute of Comp. Tech. Beijing (F. Dongrui)
  – Brasil, Univ. São Paulo (A. Pereira do Lago)
  – Argentina, Univ. Córdoba (G. Infante-Lòpez)

• National
  – Info LIFL Lille…
  – BioInfo MIG Inra Jouy, IBCP Lyon, URGI Evry…
  – Bio Institut Curie, LEPG Tours, INSERM U694 Angers…

• Numerous regional cooperations (west of France) via a consortium of more than 50 research laboratories in genomics Ouest-genopole Sea/Agro/Health (Inserm U456 Rennes, Ifremer LME Brest, Valoria Vannes, Inra génétique animale Rennes…)
Syntactical models on sequences

• **Motivation**: Allow the biologist to express, extract and check complex models on sequences, as formal languages.

1. The biologist provides a model

   Parsing genomes = String Variable Grammars + Suffix trees

   OR

2. The model has to be induced from sequences

   Grammatical Inference
Syntactical model of transposons

Genome Syntactical Analyser: STAN

http://stan.genouest.org

Nicolas J., Durand P., Ranchy G., Tempel S. and Valin A-S.

Suffix-Tree ANalyser (STAN): looking for nucleotidic and peptidic patterns in genomes.
Bioinformatics oct. 2005
Visualization and Search of structures in whole genomes

♦ CRISPR (Archae)
♦ Genomic domains (transposons)
♦ μRNA (pea aphid)

Unafold
HybridSSMIN
Speed-up X10 on GPU NVIDIA
Protein discovery in whole genomes
WAPAM: Search of weighted motifs on proteins

Sequencing (2004)

36 M STS 36 Gbp

Assembly

6-12 months

Discovery & matching of 5 degenerated patterns

63745 STS 60 Mbp

Targeted Assembly

10 hours

5568 contigs 1122 gènes OR

Expertise, Prediction of genes

Draft Assembly
July 2004

PC : 2 days
R-Disk : 15 minutes

The dog and rat olfactory receptor repertoires.
Genome Biology vol 6 N° 10 2005
Grammatical inference: characterization of membrane proteins

A family of Protein Sequences:

> AQP1_BOVIN
MASEFKKLFWRAY ....... KPK
> AQP3_MOUSE
MGROKELMNRCGE .... HLNPAVTF.... SSV
> AQP9_HUMAN
MQPEGAEKGSFQRLVKSLA ... HINPAVSLA..... SKM
> AQP4_BOVIN
M5DRPAATR4GWKCGPLCTRES ....... E01

Signature of the family:

Leave-one-out cross-validation

<table>
<thead>
<tr>
<th>Automata Size</th>
<th>Precision</th>
<th>Recall</th>
<th>F-mes.</th>
<th>Precision</th>
<th>Recall</th>
<th>F-mes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>100%</td>
<td>92%</td>
<td>0.96</td>
<td>100%</td>
<td>96%</td>
<td>0.98</td>
</tr>
<tr>
<td>40</td>
<td>100%</td>
<td>71%</td>
<td>0.83</td>
<td>100%</td>
<td>100%</td>
<td>1.00</td>
</tr>
<tr>
<td>100</td>
<td>100%</td>
<td>54%</td>
<td>0.70</td>
<td>100%</td>
<td>100%</td>
<td>1.00</td>
</tr>
<tr>
<td>130</td>
<td>100%</td>
<td>42%</td>
<td>0.59</td>
<td>100%</td>
<td>96%</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Optimization and Parallelism

- **Motivation:** Provide advanced parallel hardware and software for critical applications in genomics.

- **Volume:** Dedicated machines;

- **Complexity:**
  Parallelized Discrete optimization methods;

Data are now growing faster than the computer power.
Dedicated architectures
ReMIX project (2004-2006)

- French ministry support
  - ACI MD
- Very fast genome analysis (96 Gbp/sec)
- Parallel and dedicated system
- 4 nodes
  - 1 node = 128 Gbytes FLASH memory + FPGA resources
- Applications:
  - Content-based search (BLAST-like)
  - Whole genome comparison
  - Motif search
  - Repeat detection
  - ...

Speed-up X 50 achieved
Genomes segmentation (collab. with INRA-microbiology lab, Rennes)

Context:
- Study plasticity of bacterium genomes
- Avoid systematic sequencing of strains
- Use Long Range PCR techniques
  - Cut the genome into 10 Kbp fragments

Challenge:
- Find the best set of couple of primers to optimize the LR-PCR on the whole genome
- \( P^N \) possible solutions (\( P \sim 10, N \sim 100 \))

Results:
- Sequence segmentation modelization
- Parallel dynamic programming algorithm and GenoFrag software
- Successful biological validation

Protein threading
(collab. with INRA-MIG, Jouy en Josas)

• Given a protein sequence find the positions of each amino acid in its 3D folded shape: NP-complete problem.

• PTP has been proved to be equivalent to the augmented shortest path problem on a specific graph. New mathematical programming model has been proposed and new solver has been developed.

• PTP has been solved
Protein 900 AA, Polytope $10^{77}$ vertices, time 5' with dedicated algorithm.
Parallel version available

• Best approach to date. This new Lagrangian approach Outperforms Raptor (Univ. Waterloo M. Li, J. Xu), the best comparable approach.

R. Andonov, S. Balev, N. Yanev.
Protein Threading Problem: From Mathematical Models to Parallel Implementations,
INFORMS Journal on Computing, Special Issue on Computational Molecular Biology/Bioinformatics,
Networks Modelling

• **Motivation**: Help to understand biological networks (including genes, metabolites and signals)
  • Use qualitative measurements and incomplete information.

• **Modeling**:
  Models adapted to sparse knowledge and data: interaction graphs.
  – Knowledge: store interaction of the literature in an appropriate way.
  – Infer interaction: analysis of promotors in sequences.

• **Dynamics Analysis**:
  – Compare the prediction of the interaction graph to qualitative data.
  – Propose corrections to the model and/or data.
  – Analyse and decompose the graph to predict multistationnarity.
Large Scale modelling in Systems Biology

- Algorithmics for model reduction: Analysis of the signalization network of NF\(_k\)B by identification of critical parameters (>100 parameters) and analysis of the influence graph of E. coli (1200 nodes, 2500 edges);

- Validation and prediction of qualitative data on large transcriptional networks: complete network on E. Coli and on Yeast (2400 nodes, 4300 interactions);

- Current Challenge: Modelling the effect of a chimeric protein (EWS/FLY1) in Ewing’s tumor, ANR SITCON with Curie (100 genes, expression data, ChIP-chip and CGH)
Consistency Knowledge/ Observations

- Interaction graph:
  - vertices = genes colored by variation (positive or negative),
  - edges = regulations, colored by type (activation or inhibition)

Perturbation → equilibrium shift

Each variation must be explained by at least one influence
Prediction of the response of E. coli to starvation stress

- **RegulonDB**
  - 1500 genes;
  - 3800 regulations.

- **Literature**
  - 40 known stress variations.

- ⇒ extract minimal inconsistent subgraphs
  ⇒ error in RegulonDB
- Prediction of 26% of the consistent network
- Time: 30 s (limitation with BDD to 400 variables)
- Logical Extensions with Answer Set Programming Modelling.
Contacts

Jacques.Nicolas@irisa.fr
(leader)
Marie-Noelle.Georgeault@irisa.fr
(assistant)
Anne.Siegel@irisa.fr
(seminar)