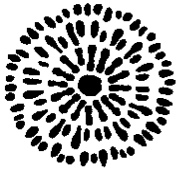
Introduction to Bioinformatics

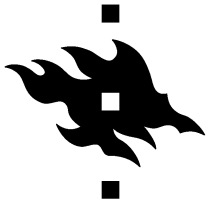
Sirkka-Liisa Varvio
sirkka-liisa.varvio@helsinki.fi

Autumn 2009, I period

www.cs.helsinki.fi/mbi/courses/09-10/itb



MBI MASTER'S DEGREE
PROGRAMME IN BIOINFORMATICS



How to enrol for the course?

- Use the registration system of the Computer Science department: https://ilmo.cs.helsinki.fi
  - You need your user account at the IT department (“cc account”, NOT “cs account” !)
  - If you cannot register yet, don’t worry: attend the lectures and exercises; just register when you are able to do so

Teachers

- Sirkka-Liisa Varvio

  Department of Mathematics and Statistics, University of Helsinki

- Veli Mäkinen

  Department of Computer Science, University of Helsinki

- Fabian Hoti

  Department of Information and Computer Science, Helsinki University of Technology

- Laura Langohr (exercises)

  Department of Computer Science, University of Helsinki

Lectures and exercises

- Lectures:

  Tuesday and Thursday 14^{15} - 16 Exactum D122

  - *Lectures start Tuesday 8. September*

- Exercises:

  Group 1 Thursday 12^{15} - 14 Exactum BK106
  Group 2 Thursday 16^{15} - 18 Exactum BK 106
  Group 3 if needed

  - *Exercises start Thursday 17. September*

Status & Prerequisites

- Advanced level course, suitable also for intermediate studies
- 4 credits (2 credits without exercises)

- Prerequisites
  - *Basic mathematics and statistics skills, probability calculus*
  - *Familiarity with computers*
  - *Basic programming skills recommended*
  - *No biology background required*

Course contents

- What is bioinformatics?
- Basics on bases: A-G-T-C as words
- Biological basics and bioinformatic challenges
- - “ -
- Sequence alignment (and assembly)
- - “ -
- - “ -
- - “ -
- Gene expression analysis
- Phylogenetic trees, inferring the past
- Population genomics, genetic variation, haplotype analysis
- Comparative genomics

How to pass the course?

- Recommended method:
  - Attend the lectures (not obligatory though)
  - Do the exercises
  - Take the course exam Wednesday 21. October, 16.00 – 19.00
- Or:
  - Take a separate exam
    - See the websites of Department of Computer Science for separate examinations

How to pass the course?

- Exercises give you max. 12 points
  - 0% completed assignments gives you 0 points, 80% gives 12 points, the rest by linear interpolation
  - "A completed assignment" means that
    - You are willing to present your solution in the exercise session and
    - you return notes by e-mail to Laura Langohr describing the main phases you took to solve the assignment
    - Return notes at latest on Thursdays 12^{15}
- Course exam gives you max. 48 points

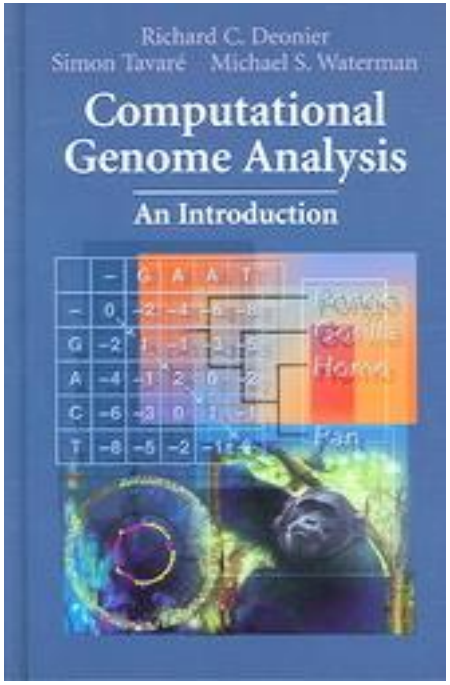
How to pass the course?

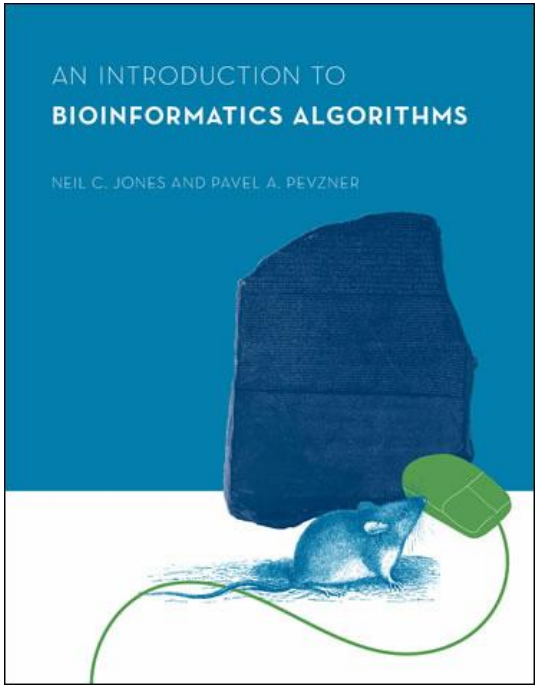
- Grading: on the scale 0-5
  - To get the lowest passing grade 1, you need to get at least 30 points out of 60 maximum
- Course exam: Wednesday 21. October 16.**00**-19.00 Exactum A111

- If you take the first separate exam, the best of the following options will be considered:
  - Exam gives you 48 points, exercises 12 points
  - Exam gives you 60 points
- In second and subsequent separate exams, only the 60 point option is in use

Literature

- **Deonier, Tavaré, Waterman:** ***Computational Genome Analysis, an Introduction. Springer, 2005***
- **Jones, Pevzner:** ***An Introduction to Bioinformatics Algorithms. MIT Press, 2004***
- **You are not supposed to read these books to the examination which is based on lectures and exercises.**
- **Lectures do not literally obey book material and may include other (for example: more recent) material, depending on the topic.**





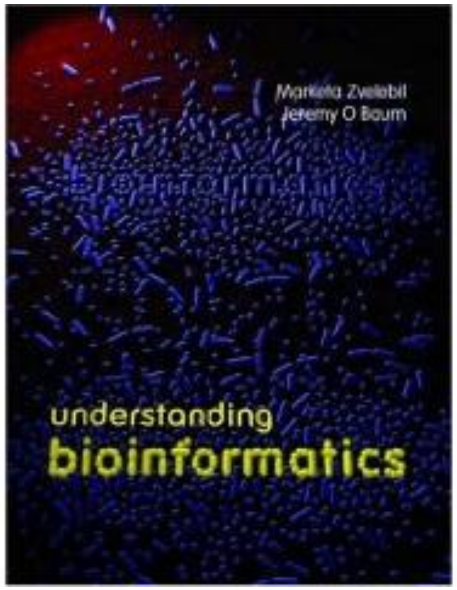
Additional literature

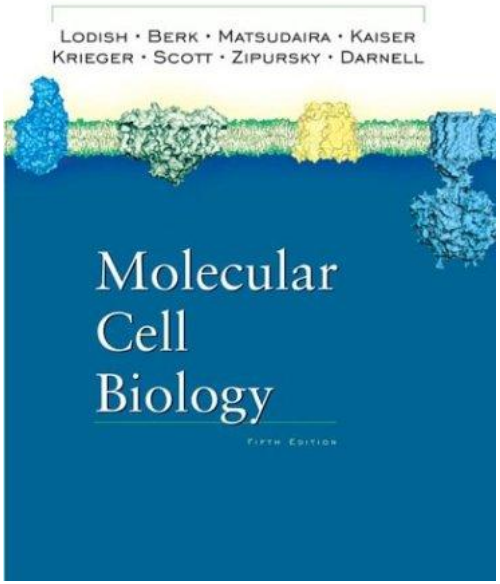
- **Many of the picture slides in the lectures are taken from:**

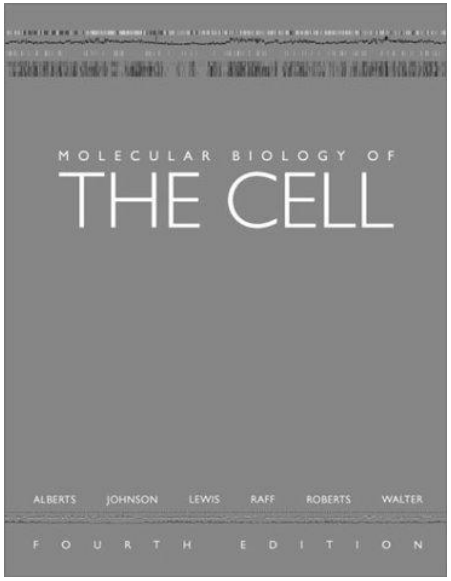
Zvelebid & Baum,
Understanding Bioinformatics, Garland Science, 2007.

- **Basic books about molecular biology:**

Alberts et al.: ***Molecular biology of the cell***
Lodish et al.: ***Molecular cell biology***







Master's Degree Programme in Bioinformatics (MBI)

- in a nutshell

- Two-year international MSc programme
- Admission for 2010-2011 in January 2010
  - You need to have your Bachelor’s degree ready by August 2010

MBI programme organizers

Department of Computer Science,
Department of Mathematics and Statistics
Faculty of Science, Kumpula Campus,
University of Helsinki

Laboratory of Computer and Information Science, Laboratory of CS and Engineering, *TKK*

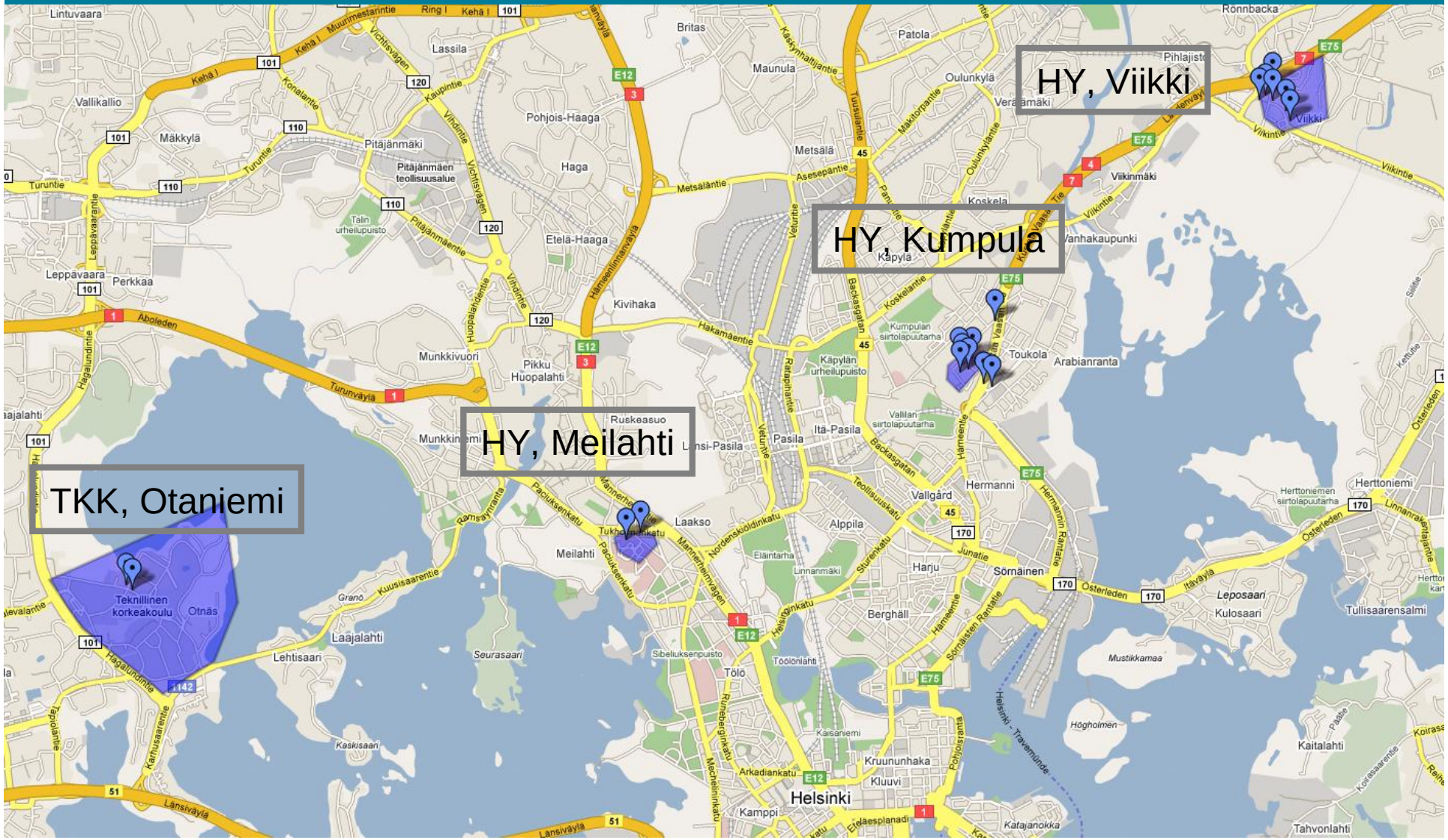
……are responsible for bioinformatics major subject studies and computer science, mathematics and statistics minor subject studies

Faculty of Medicine, Meilahti Campus,
University of Helsinki

Faculty of Biosciences
Faculty of Agriculture and Forestry
Viikki Campus, University of Helsinki

…. organize tailor-made and other biology courses for minor subject studies

Four MBI campuses



What is bioinformatics?

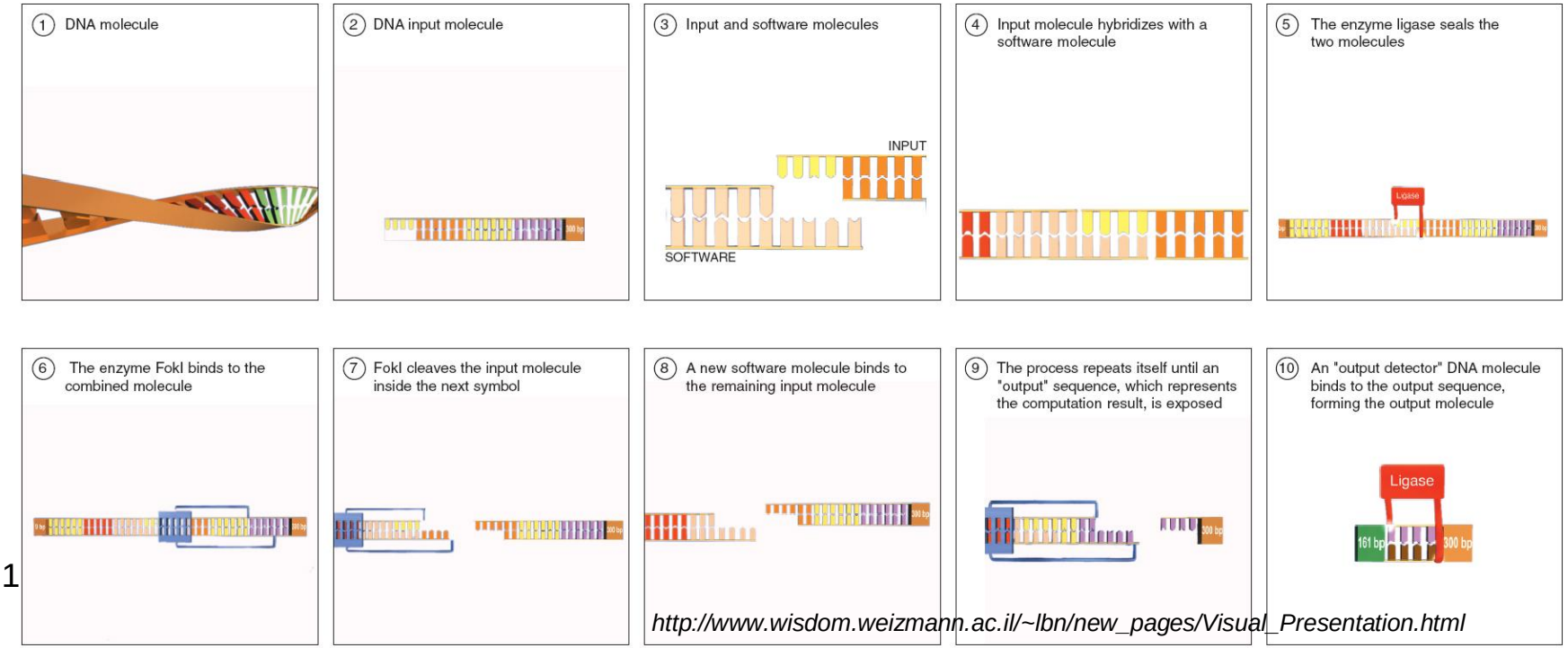
- Bioinformatics, *n.* The science of information and information flow in biological systems, esp. of the use of computational methods in genetics and genomics. (Oxford English Dictionary)

- "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.“

- "I do not think all biological computing is bioinformatics, e.g. mathematical modelling is not bioinformatics, even when connected with biology-related problems. In my opinion, bioinformatics has to do with management and the subsequent use of biological information, particular

What is *not* bioinformatics?

- Biologically-inspired computation, e.g., genetic algorithms and neural networks
- However, application of neural networks to solve some biological problem, could be called bioinformatics
- What about DNA computing?



http://www.wisdom.weizmann.ac.il/~lbn/new_pages/Visual_Presentation.html

Computational biology

- **Application of computing to biology (broad definition)**
- **Often used interchangeably with bioinformatics**
- **Or: *Biology* that is done with computational means**

Mathematical biology

- Mathematical biology “tackles biological problems, but the methods it uses to tackle them need not be numerical and need not be implemented in software or hardware.”

Alan Turing

THE CHEMICAL BASIS OF MORPHOGENESIS

By A. M. TURING, F.R.S. *University of Manchester*

(*Received* 9 *November* 1951—*Revised* 15 *March* 1952)

It is suggested that a system of chemical substances, called morphogens, reacting together and diffusing through a tissue, is adequate to account for the main phenomena of morphogenesis. Such a system, although it may originally be quite homogeneous, may later develop a pattern or structure due to an instability of the homogeneous equilibrium, which is triggered off by random disturbances. Such reaction-diffusion systems are considered in some detail in the case of an isolated ring of cells, a mathematically convenient, though biologically unusual system. The investigation is chiefly concerned with the onset of instability. It is found that there are six essentially different forms which this may take. In the most interesting form stationary waves appear on the ring. It is suggested that this might account, for instance, for the tentacle patterns on *Hydra* and for whorled leaves. A system of reactions and diffusion on a sphere is also considered. Such a system appears to account for gastrulation. Another reaction system in two dimensions gives rise to patterns reminiscent of dappling. It is also suggested that stationary waves in two dimensions could account for the phenomena of phyllotaxis.

The purpose of this paper is to discuss a possible mechanism by which the genes of a zygote may determine the anatomical structure of the resulting organism. The theory does not make any new hypotheses; it merely suggests that certain well-known physical laws are sufficient to account for many of the facts. The full understanding of the paper requires a good knowledge of mathematics, some biology, and some elementary chemistry. Since readers cannot be expected to be experts in all of these subjects, a number of elementary facts are explained, which can be found in text-books, but whose omission would make the paper difficult reading.

1. A model of the embryo. Morphogens

In this section a mathematical model of the growing embryo will be described. This model will be a simplification and an idealization, and consequently a falsification. It is to be hoped that the features retained for discussion are those of greatest importance in the present state of knowledge.

The model takes two slightly different forms. In one of them the cell theory is recognized but the cells are idealized into geometrical points. In the other the matter of the organism is imagined as continuously distributed. The cells are not, however, completely ignored, for various physical and physico-chemical characteristics of the matter as a whole are assumed to have values appropriate to the cellular matter.

With either of the models one proceeds as with a physical theory and defines an entity called ‘the state of the system’. One then describes how that state is to be determined from the state at a moment very shortly before. With either model the description of the state consists of two parts, the mechanical and the chemical. The mechanical part of the state describes the positions, masses, velocities and elastic properties of the cells, and the forces between them. In the continuous form of the theory essentially the same information is given in the form of the stress, velocity, density and elasticity of the matter. The chemical part of the state is given (in the cell form of theory) as the chemical composition of each separate cell; the diffusibility of each substance between each two adjacent cells must also

Vol. 237. B. 641. (Price 8s.) 5 [Published 14 August 1952

Turing on biological complexity

- “It must be admitted that the biological examples which it has been possible to give in the present paper are very limited.

  This can be ascribed quite simply to the fact that biological phenomena are usually very complicated. Taking this in combination with the relatively elementary mathematics used in this paper one could hardly expect to find that many observed biological phenomena would be covered.

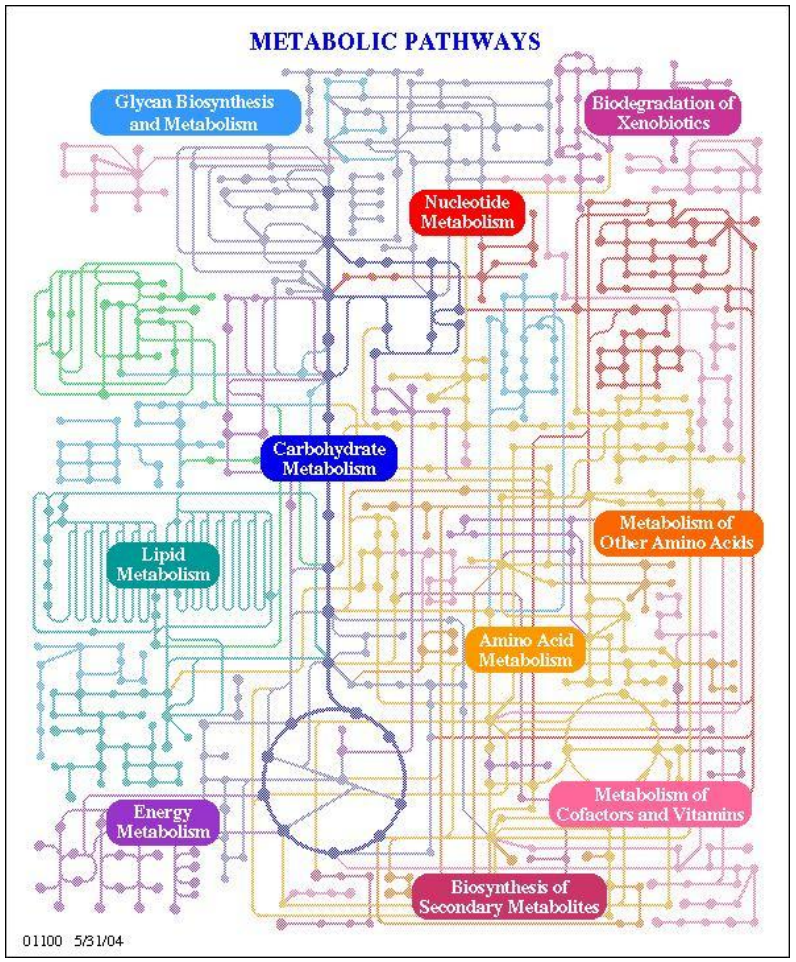
  It is thought, however, that the imaginary biological systems which have been treated, and the principles which have been discussed, should be of some help in interpreting real biological forms.”

  – Alan Turing, The Chemical Basis of Morphogenesis, 1952

Related concepts

- Systems biology
  - “Biology of networks”
  - Integrating different levels of information to understand how biological systems work

- Computational systems biology



Overview of metabolic pathways in KEGG database, www.genome.jp/kegg/

Why is bioinformatics important?

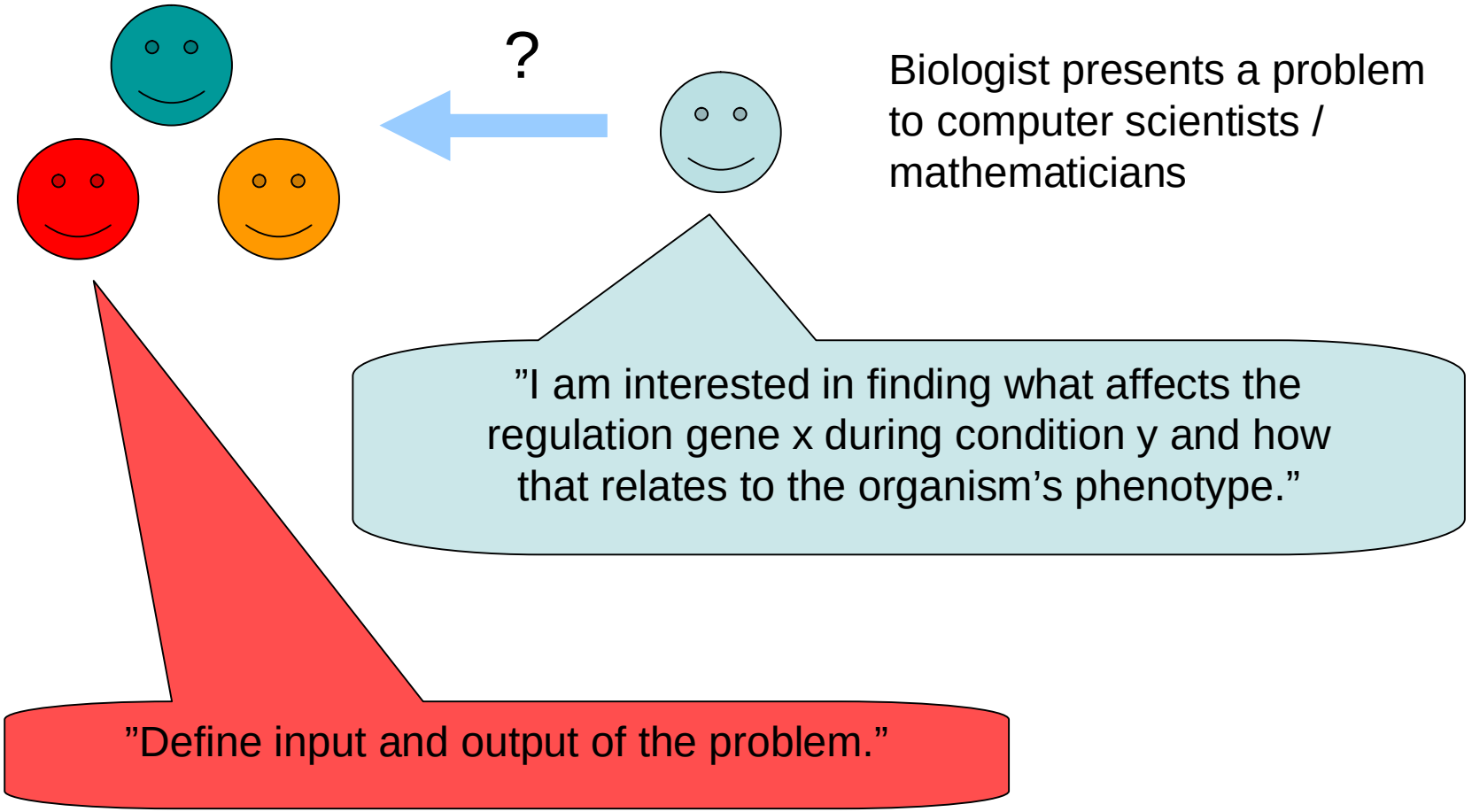
- New measurement techniques produce huge quantities of biological data
  - Advanced data analysis methods are needed to make sense of the data
  - Typical data sources produce noisy data with a lot of missing values
- Paradigm shift in biology to utilise bioinformatics in research

Bioinformatician's skill set

- Statistics, data analysis methods
  - Lots of data
  - High noise levels, missing values
  - #attributes >> #data points
- Programming languages
  - Scripting languages: Python, Perl, Ruby, …
  - Extensive use of text file formats: need parsers
  - Integration of both data and tools
- Data structures, databases
- Modelling
  - Discrete vs continuous domains
  - -> Systems biology
- Scientific computation packages

From: Esa Pitkänen

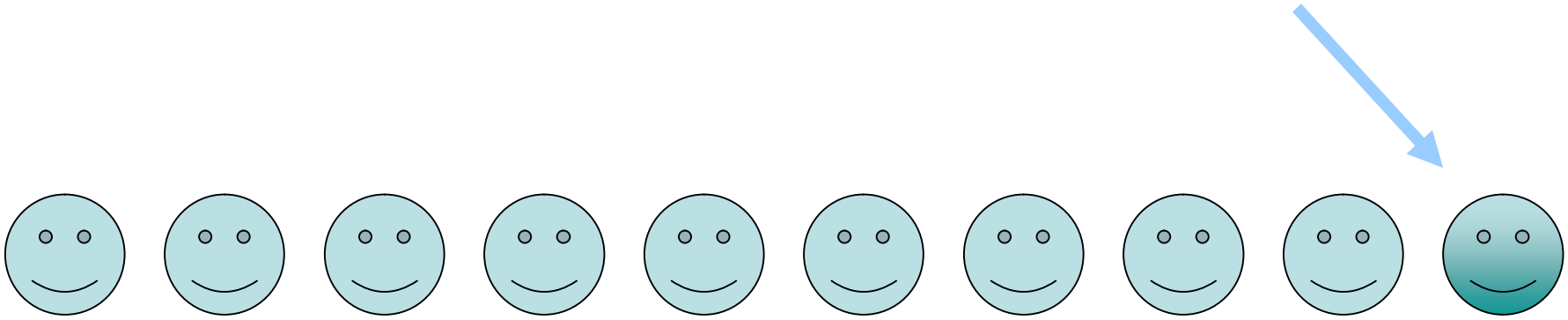
Communication skills: case 1



From: Esa Pitkänen

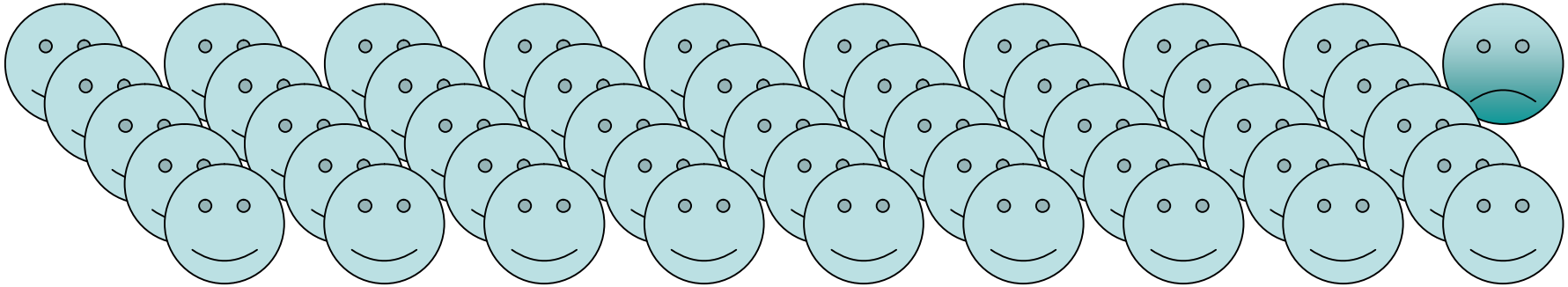
Communication skills: case 2

Bioinformatician is a part of a group that consists mostly of biologists.



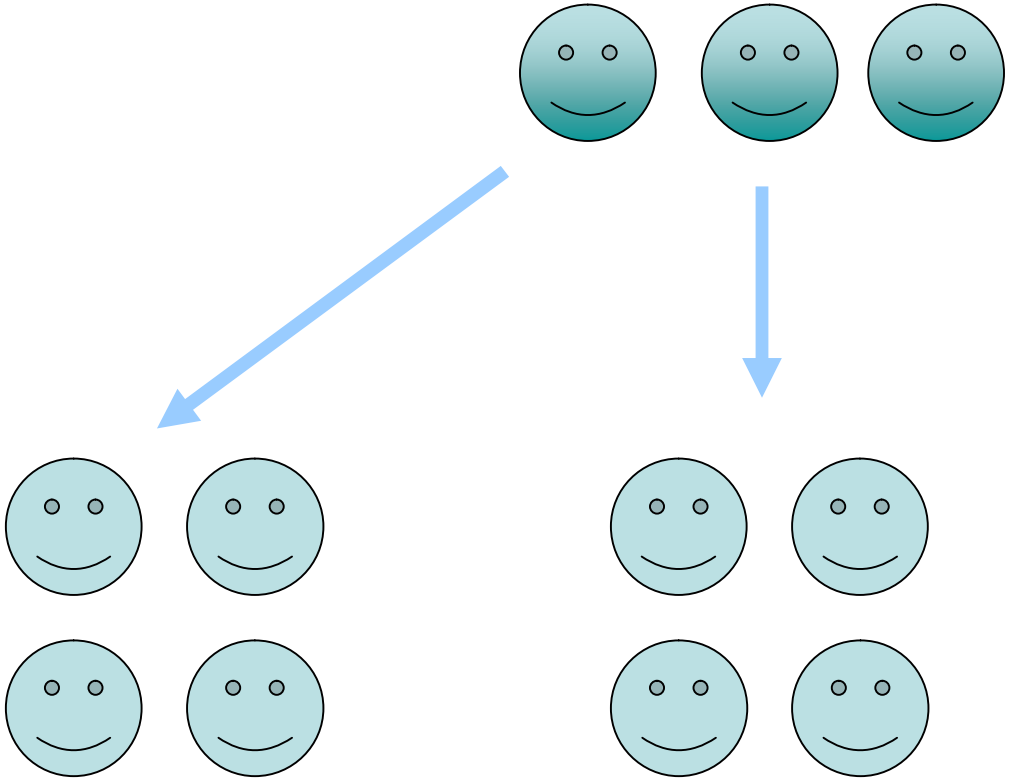
From: Esa Pitkänen

Communication skills: case 2



...biologist/bioinformatician ratio is important!

Communication skills: case 3



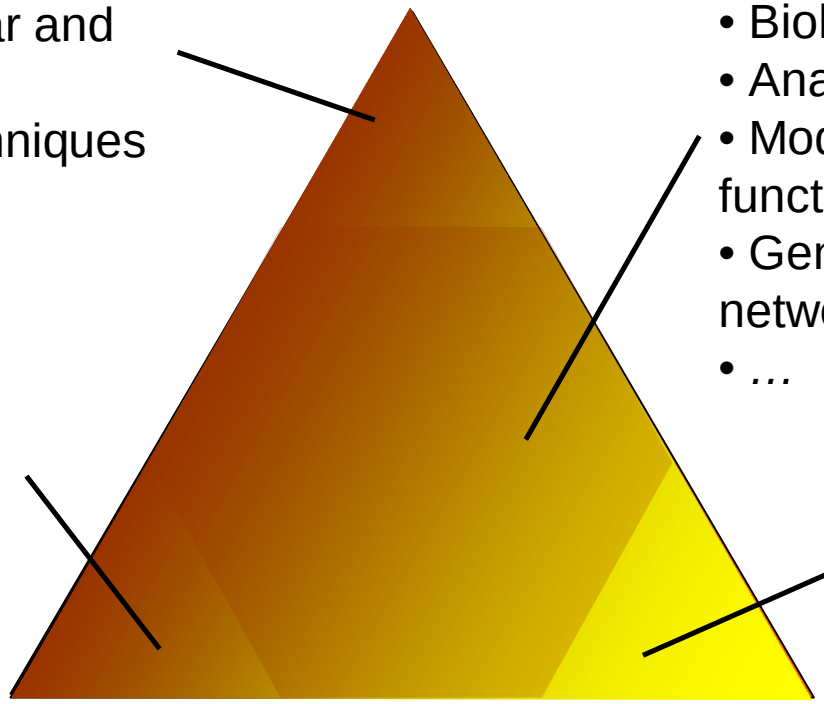
A group of bioinformaticians offers their services to more than one group

From: Juho Rousu

Bioinformatician's skill set

Biology & Medicine
- Basics in molecular and cell biology
- Measurement techniques

Computer Science
- Programming
- Databases
- Algorithmics



Bioinformatics
- Biological sequence analysis
- Biological databases
- Analysis of gene expression
- Modeling protein structure and function
- Gene, protein and metabolic networks
- …

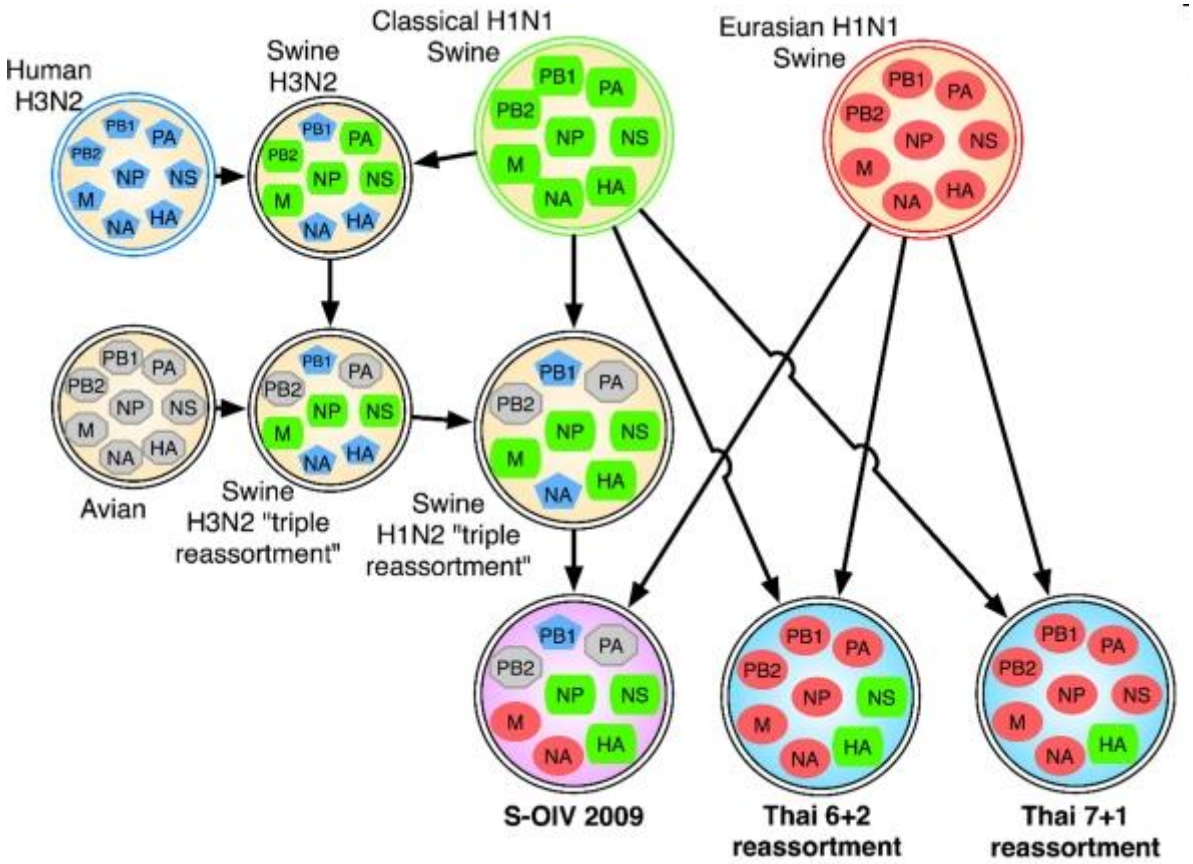
Mathematics and statistics
- Calculus
- Probability calculus
- Linear algebra

Where would you be in this triangle?

An example of importance of bioinformatics:

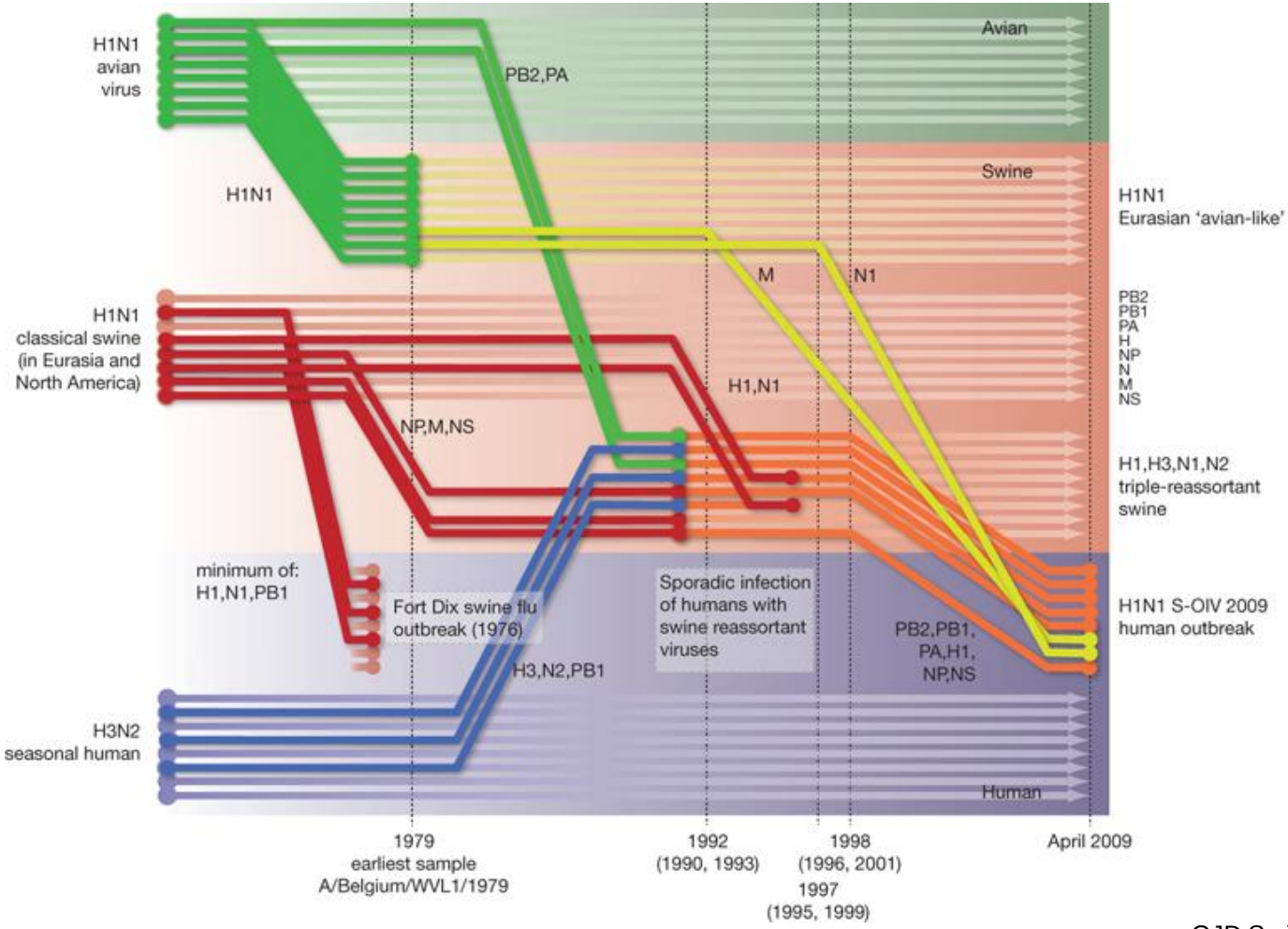
UNDERSTANDING THE SWINE FLU, H1N1, REQUIRES BIOINFORMATICS

Reassortment history of the 2009 H1N1 outbreak strain and the Thai reassortants



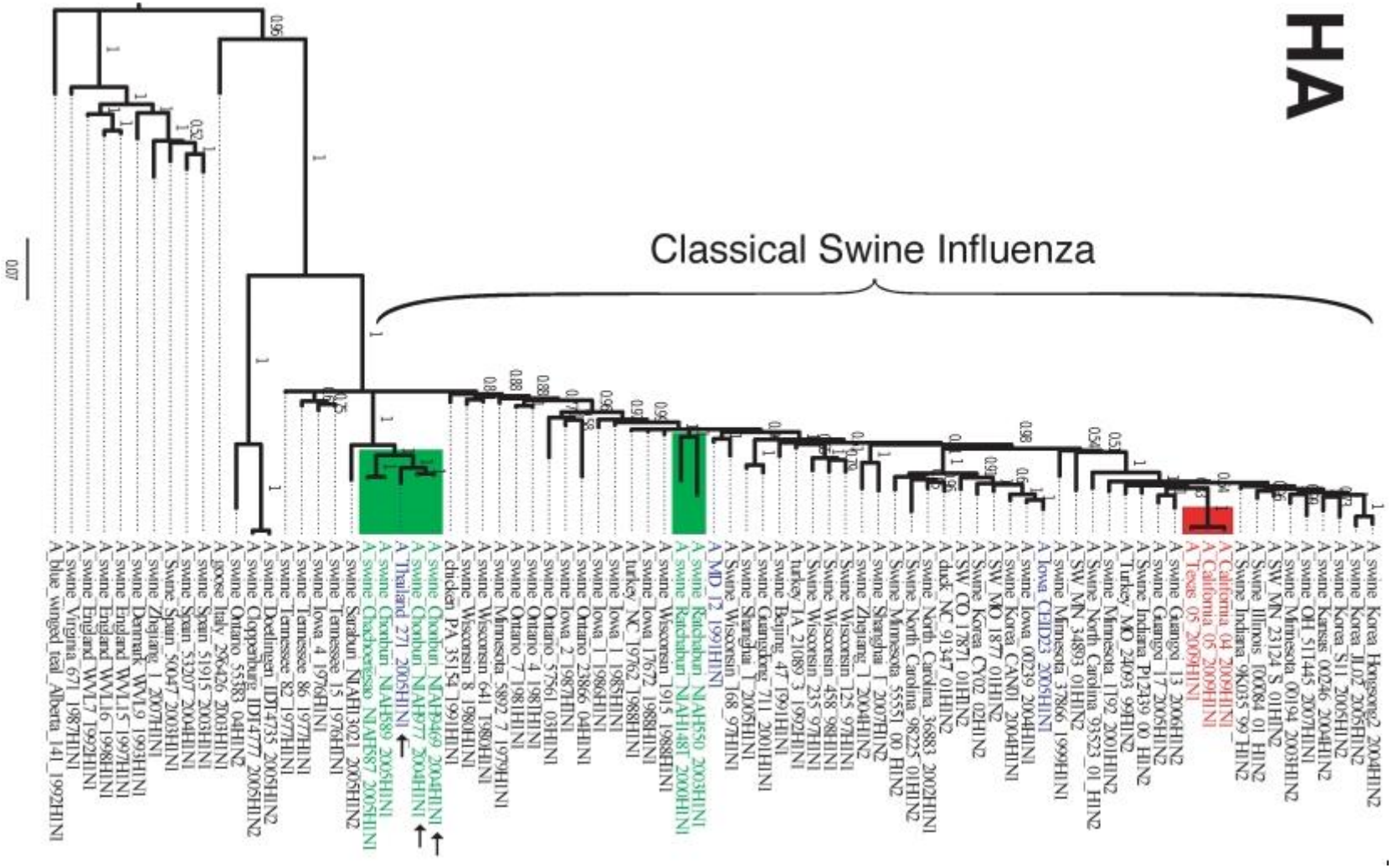
PLoS One. 2009; 4(7): e6402.
Published online 2009 July 28.
doi:
10.1371/journal.pone.0006402.

Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic



Shaded boxes represent host species; avian (green), swine (red) and human (grey). Coloured lines represent interspecies-transmission pathways of influenza genes. The eight genomic segments are represented as parallel lines in descending order of size. Dates marked with dashed vertical lines on 'elbows' indicate the mean time of divergence of the S-OIV genes from corresponding virus lineages. Reassortment events not involved with the emergence of human disease are omitted. Fort Dix refers to the last major outbreak of S-OIV in humans. The first triple-reassortant swine viruses were detected in 1998, but to improve clarity the origin of this lineage is placed earlier.

GJD Smith *et al. Nature* **459**, 1122-1125 (2009)
doi:10.1038/nature08182



PLoS One. 2009; 4(7): e6402.
Published online 2009 July 28. doi:
10.1371/journal.pone.0006402.