BioInfo Paper — Topic II

Falk Schubert

In the beginning, the new sequence data has to be stored in some database system. This helps other people to compare their sequences and improves access of the genomic data. New extracted data – either through experiments or through database research – will also be stored in a database.

One of first questions that comes up when exploring a new genome of an organism is: What organisms are closely related to the new one and what data do I already have? The more genomic data is submitted to large database the harder it is to keep an overview about work that has been done. To avoid duplicated work and to benefit from synergy effects through comparative analysis, answering this question can help a lot.

Since it is not always clear how an organism should be classified correctly from the biological features that makes this organism novel and to get a measure of what "closely related" means, a homology analysis can be done. Some genes – the corresponding sequences can be extracted from the ORFs – can be used for alignment against different organisms. For simplification the considered organims should be limited to the suspected ones. Programs like *BLAST* can be used for this. If we find some similar sequences in other organisms the degree of similarity might help to relate them from an evolutionary point of view.

To find more about the metabolics of this organism, some present proteins need to be identified. Using protein localization and mass spectrometry would give a little insight what and where proteins are present . Also mutations can be induced to create knock-out versions of the organism to see which genes are essential. Any unknown proteins could then be target for structure discovery. Either through x-ray-crystallography or through mapping the protein sequence to similar proteins of which the structur is known.

To find expressed genes, we can sequence the mRNA of the organism. Once we have identified the expressed genes, one could create a regulatory network by finding transcription factors to each of these genes. A possible way to do this, is to compare genes to homologous ones and search for transcription factors of the homologous genes within this new organism. Furthermore global protein-protein interactions could be found by using methods like *Two Hybrid* or *Protein Chips*. With the interaction data an initial metabolic network could be created. Again, incorporating data from related organisms can be used to complete missing links. Besides this more sophisticated mathematical calculations like clustering on graphs can performed to get predictions about possible interactions.

It is clear that most of this work requires a lot of work and time. Some analysis can be performed in parallel, like identification of expressed genes and proteins. Other problems like the construction of the networks depend on data that has to retrieved first. Hence not all of the work can be at once but it is surely possible to at least start with most the work and include the new data as it is produced.

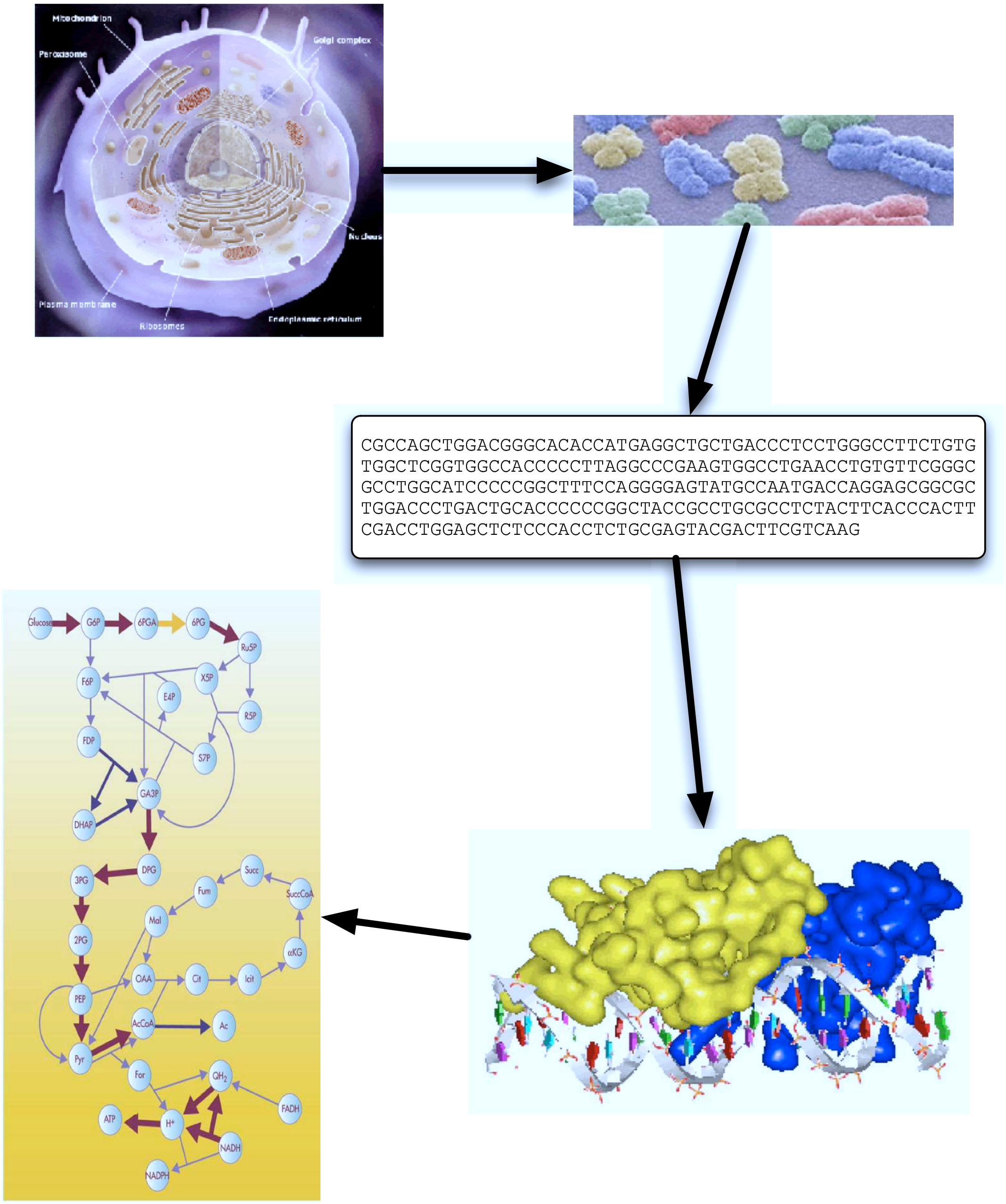
Mitochondrion
Peroxisome
Golgi complex
Nucleus
Plasma membrane
Ribosomes
Endoplasmic reticulum
CGCCAGCTGGACGGGCACACCATGAGGCTGCTGACCCTCCTGGGCCTTCTGTG
TGGCTCGGTGGCCACCCCCTTAGGCCCGAAGTGGCCTGAACCTGTGTTCGGGC
GCCTGGCATCCCCCGGCTTTCCAGGGGAGTATGCCAATGACCAGGAGCGGCGC
TGGACCCTGACTGCACCCCCCGGCTACCGCCTGCGCCTCTACTTCACCCACTT
CGACCTGGAGCTCTCCCACCTCTGCGAGTACGACTTCGTCAAG
Glucose
G6P
6PGA
6PG
Ru5P
F6P
X5P
E4P
R5P
FDP
S7P
GA3P
DHAP
DPG
3PG
Succ
Fum
SuccCoA
Mal
2PG
αKG
OAA
Cit
Icit
PEP
AcCoA
Ac
Pyr
For
QH$_2$
FADH
ATP
H$^+$
NADH
NADPH

Fig. 1: overview of genomics project: first extract dna from cell – then sequence genome – protein identification – metabolic networks