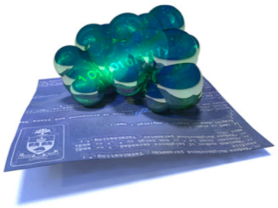
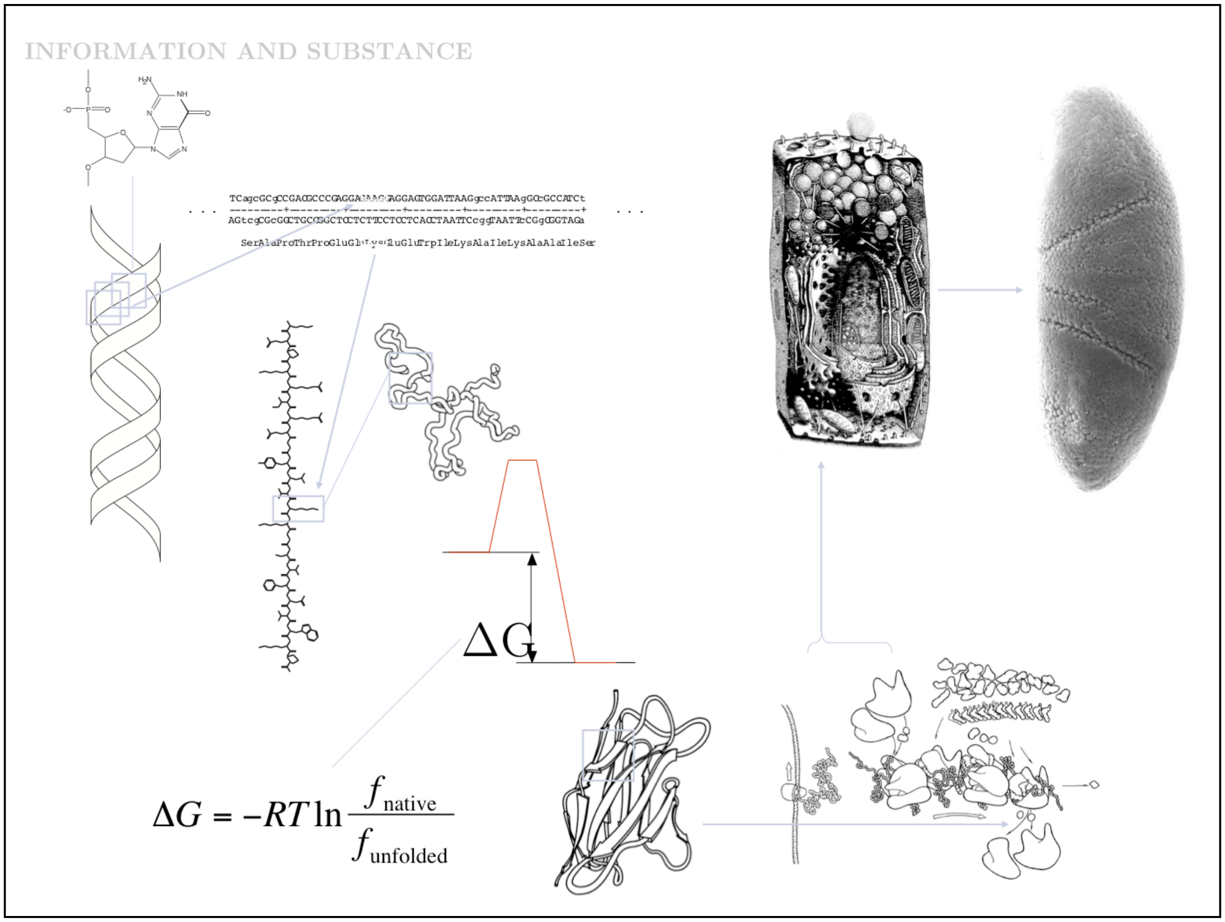
A Bioinformatics Course

Bioinformatics Concepts



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto


INFORMATION AND SUBSTANCE
H_2N
NH
O
N
. . . TCagcGCgCCGACGCCCGAGGAGAAGGAGGAGTGGATTAAGgccATTAAgGCcGCCATCt . . .
AGtcgCGcGGCTGCGGCTCCTCTTCCTCCTCACCTAATTCcggTAATTcCGgCGGTAGa
SerAlaProThrProGluGluLysGluGluTrpIleLysAlaIleLysAlaAlaIleSer
ΔG
$\Delta G = -RT \ln \frac{f_{\text{native}}}{f_{\text{unfolded}}}$

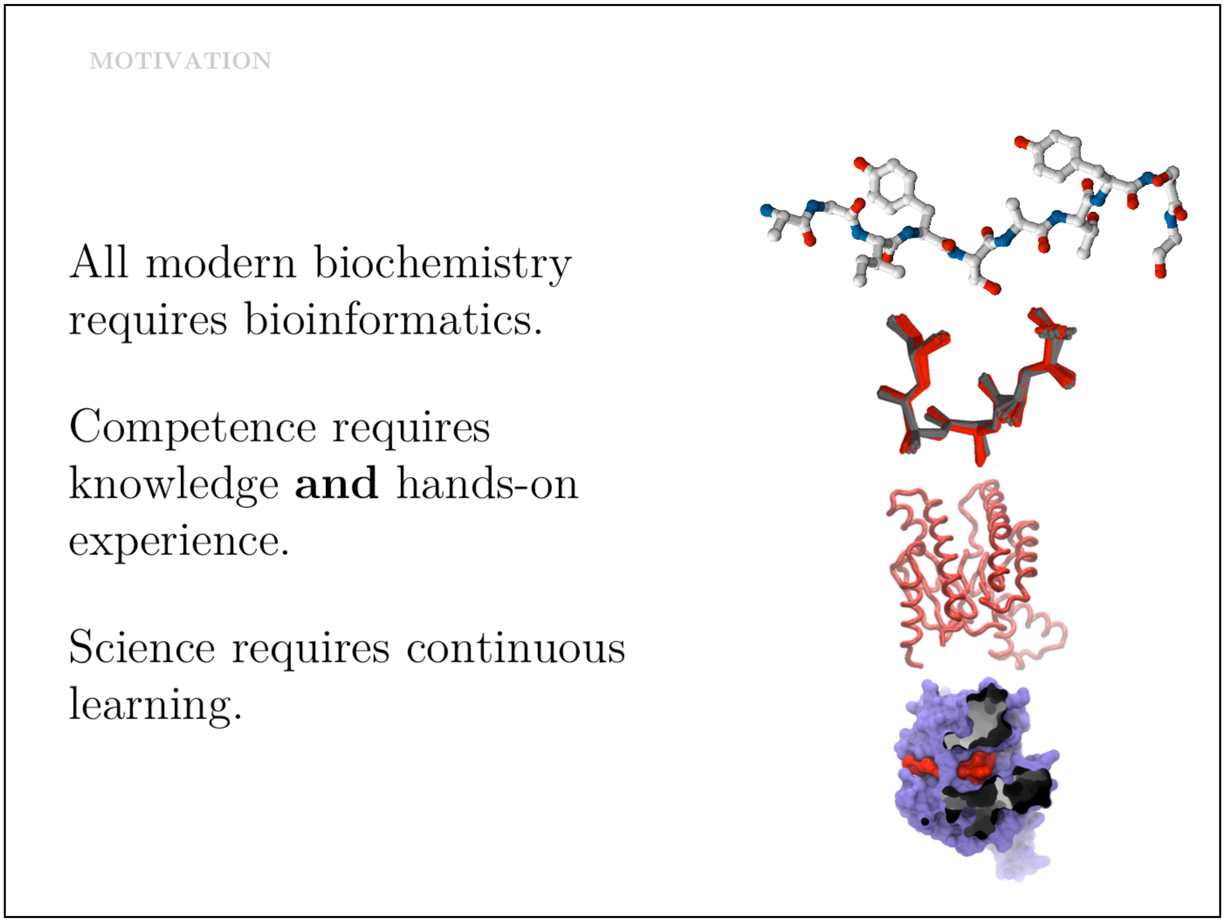
Molecular biology is an information science, just as much as a molecular science. The great experimental and conceptual advances of the 20th century have documented an unbroken flow of information: from the storage of information in the non-random sequence of nucleotide heterocopolymers, to the self-organized acquisition of structure and function in proteins, which in turn provides a selective advantage for evolution and thus influences the information that is stored in the genome.

The abstractions and models that focus on inheritable information, rather than on the details of its representation, have proven to be remarkably powerful in explaining the basic features of life, such as robust self-organization and the process–and consequences–of evolution.

In the genomic era, the relationship between information and molecule becomes ever more apparent.

In principle, all the information that is required to specify an organism is contained in its genome. This is trivially implied by the successful whole-genome synthesis experiments. The genome itself can be fully sequenced, therefore the information it contains is easily accessible to us. However, the expression of the information is organized in a hierarchical fashion, in **complex**, interacting subsystems. Knowledge of a DNA sequence does not (yet) allow us to predict the protein's structure. Knowledge of a protein's structure does not (yet) allow us to predict its interactions and assembly to molecular "machines". Knowledge of these complexes does not (yet) allow us to piece together their functional connections, as they build up the metabolic or regulatory systems, or the structural framework of a cell.

At each level, incomplete information prevents us from predicting the next-higher level of organization from its components. The sheer volume of data is a comparatively minor obstacle.



MOTIVATION

All modern biochemistry requires bioinformatics.

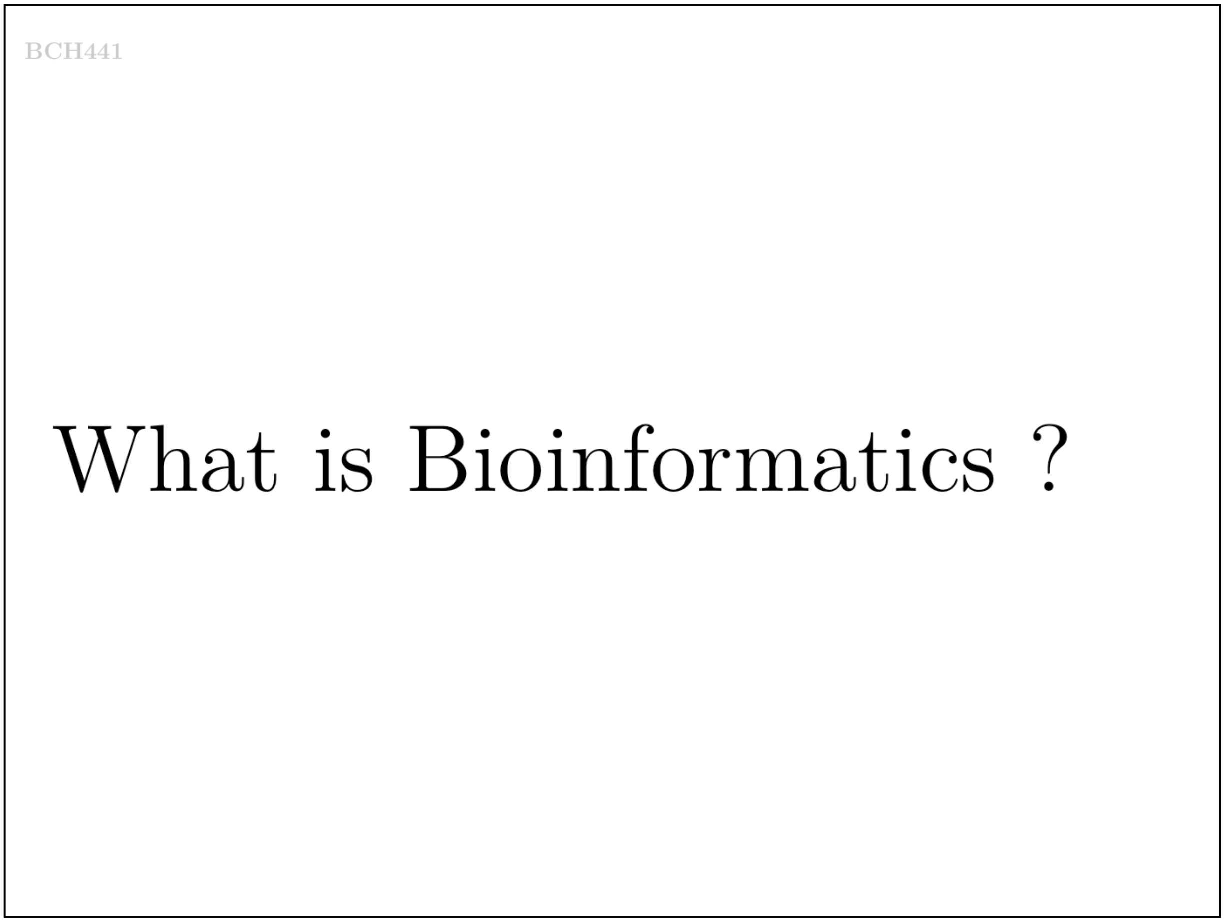
Competence requires knowledge **and** hands-on experience.

Science requires continuous learning.

Knowledge in science is always dynamic. The field changes, and the methods change too. When we solve a problem, we move on to the next one. Our targets are moving.

The information of how to work as a cutting-edge bioinformatician may half a half-life of only two years or so.

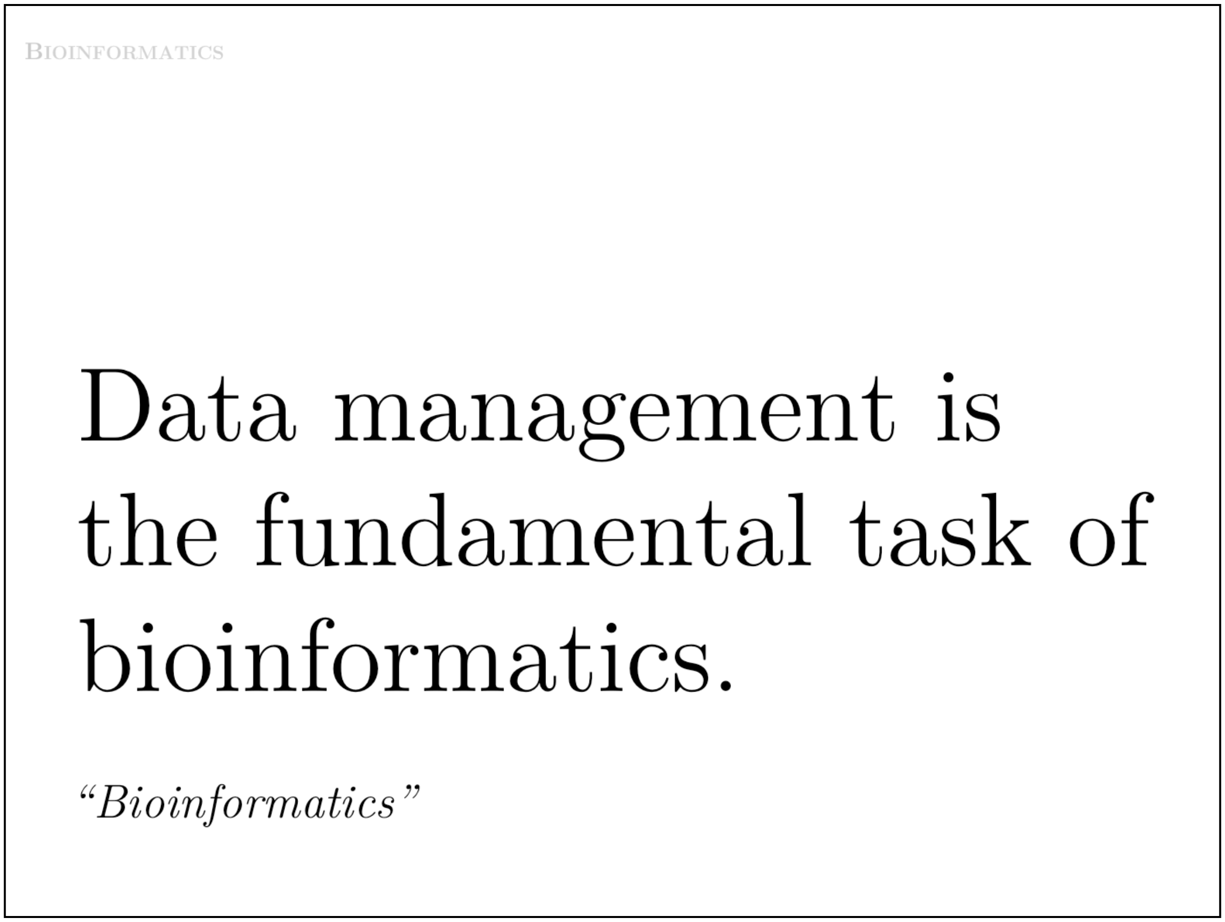
Learning **how** to learn is as important as learning any particular fact.



BCH441

What is Bioinformatics ?

As I see it, Bioinformatics is a set of paradigms that can be roughly said to span two poles:



BIOINFORMATICS

Data management is the fundamental task of bioinformatics.

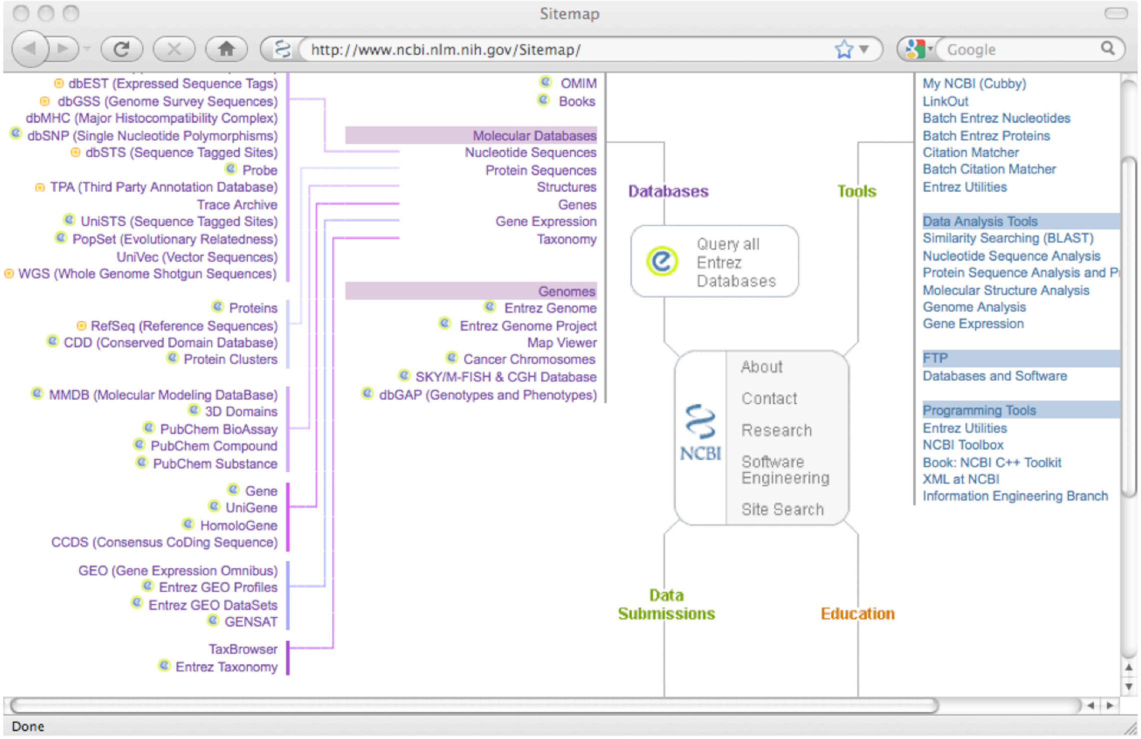
"Bioinformatics"

On one hand, if we look at the practice of bioinformatics, with its many on-line databases that need to be curated, integrated, kept consistent and made queryable, we can conclude that biological data management is what bioinformatics is all about.

Important examples of such data resources include ...

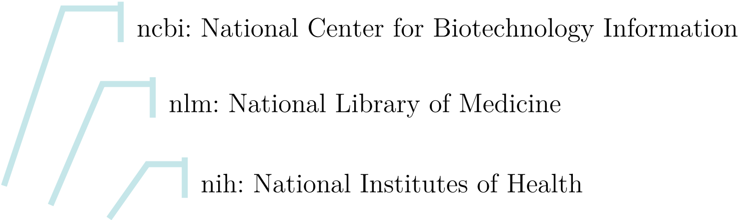
NCBI

http://www.ncbi.nlm.nih.gov


Sitemap
http://www.ncbi.nlm.nih.gov/Sitemap/
Google
dbEST (Expressed Sequence Tags)
dbGSS (Genome Survey Sequences)
dbMHC (Major Histocompatibility Complex)
dbSNP (Single Nucleotide Polymorphisms)
dbSTS (Sequence Tagged Sites)
Probe
TPA (Third Party Annotation Database)
Trace Archive
UniSTS (Sequence Tagged Sites)
PopSet (Evolutionary Relatedness)
UniVec (Vector Sequences)
WGS (Whole Genome Shotgun Sequences)
Proteins
RefSeq (Reference Sequences)
CDD (Conserved Domain Database)
Protein Clusters
MMDB (Molecular Modeling DataBase)
3D Domains
PubChem BioAssay
PubChem Compound
PubChem Substance
Gene
UniGene
HomoloGene
CCDS (Consensus CoDing Sequence)
GEO (Gene Expression Omnibus)
Entrez GEO Profiles
Entrez GEO DataSets
GENSAT
TaxBrowser
Entrez Taxonomy
OMIM
Books
Molecular Databases
Nucleotide Sequences
Protein Sequences
Structures
Genes
Gene Expression
Taxonomy
Genomes
Entrez Genome
Entrez Genome Project
Map Viewer
Cancer Chromosomes
SKY/M-FISH & CGH Database
dbGAP (Genotypes and Phenotypes)
Databases
Tools
Query all Entrez Databases
About
Contact
Research
Software Engineering
Site Search
NCBI
Data Submissions
Education
My NCBI (Cubby)
LinkOut
Batch Entrez Nucleotides
Batch Entrez Proteins
Citation Matcher
Batch Citation Matcher
Entrez Utilities
Data Analysis Tools
Similarity Searching (BLAST)
Nucleotide Sequence Analysis
Protein Sequence Analysis and P
Molecular Structure Analysis
Genome Analysis
Gene Expression
FTP
Databases and Software
Programming Tools
Entrez Utilities
NCBI Toolbox
Book: NCBI C++ Toolkit
XML at NCBI
Information Engineering Branch
Done

... the US **NCBI** (National Center for Biotechnology Information)] is one of the world's major centres for molecular data, especially sequence and genome data.

This map of the integrated NCBI resources can serve as a representative of the multitude of data and services that are availbale for molecular biology. But these tools need to be constructed, populated and maintained, and used judiciously.

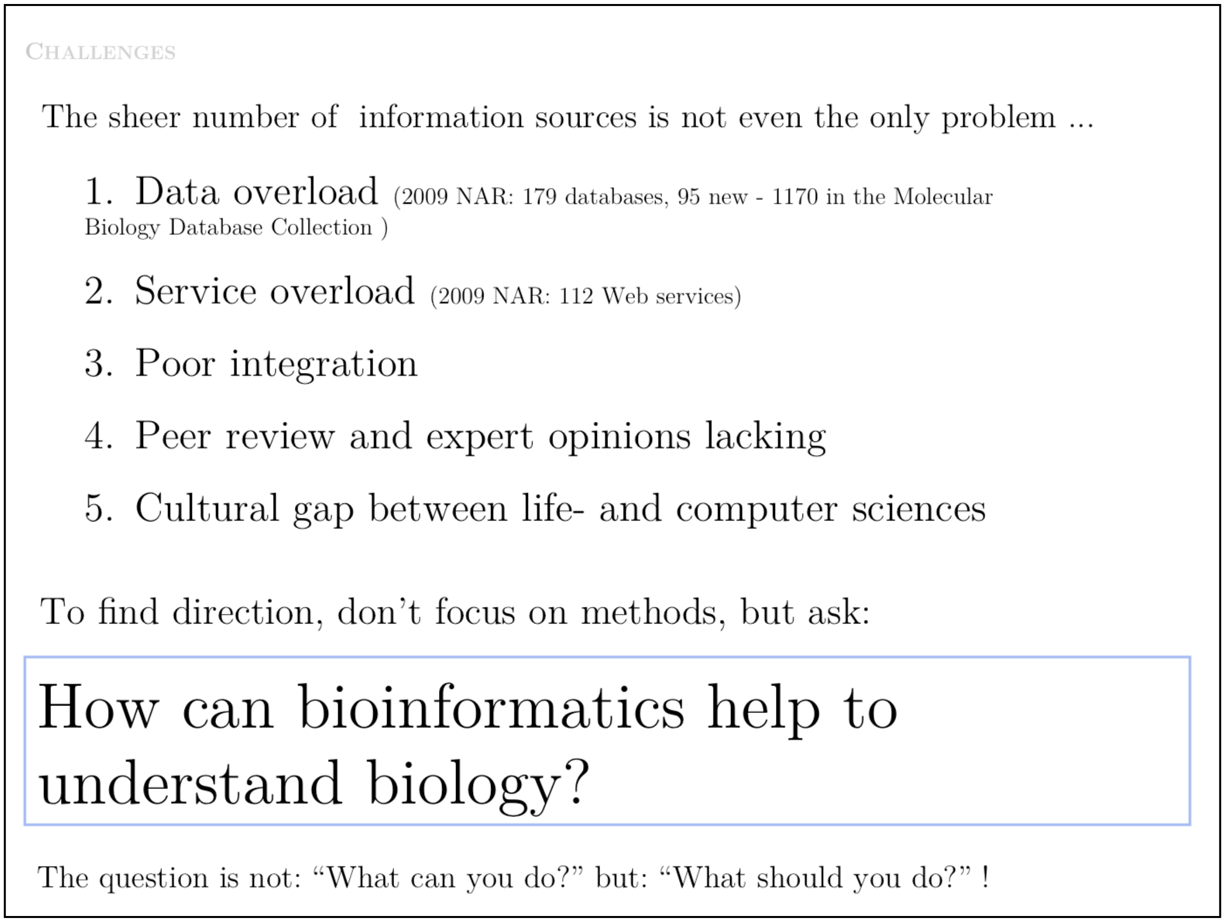


ncbi: National Center for Biotechnology Information

nlm: National Library of Medicine

nih: National Institutes of Health

(http://www.ncbi.nlm.nih.gov)



Challenges

The sheer number of information sources is not even the only problem ...

1. Data overload (2009 NAR: 179 databases, 95 new - 1170 in the Molecular Biology Database Collection)
2. Service overload (2009 NAR: 112 Web services)
3. Poor integration
4. Peer review and expert opinions lacking
5. Cultural gap between life- and computer sciences

To find direction, don’t focus on methods, but ask:

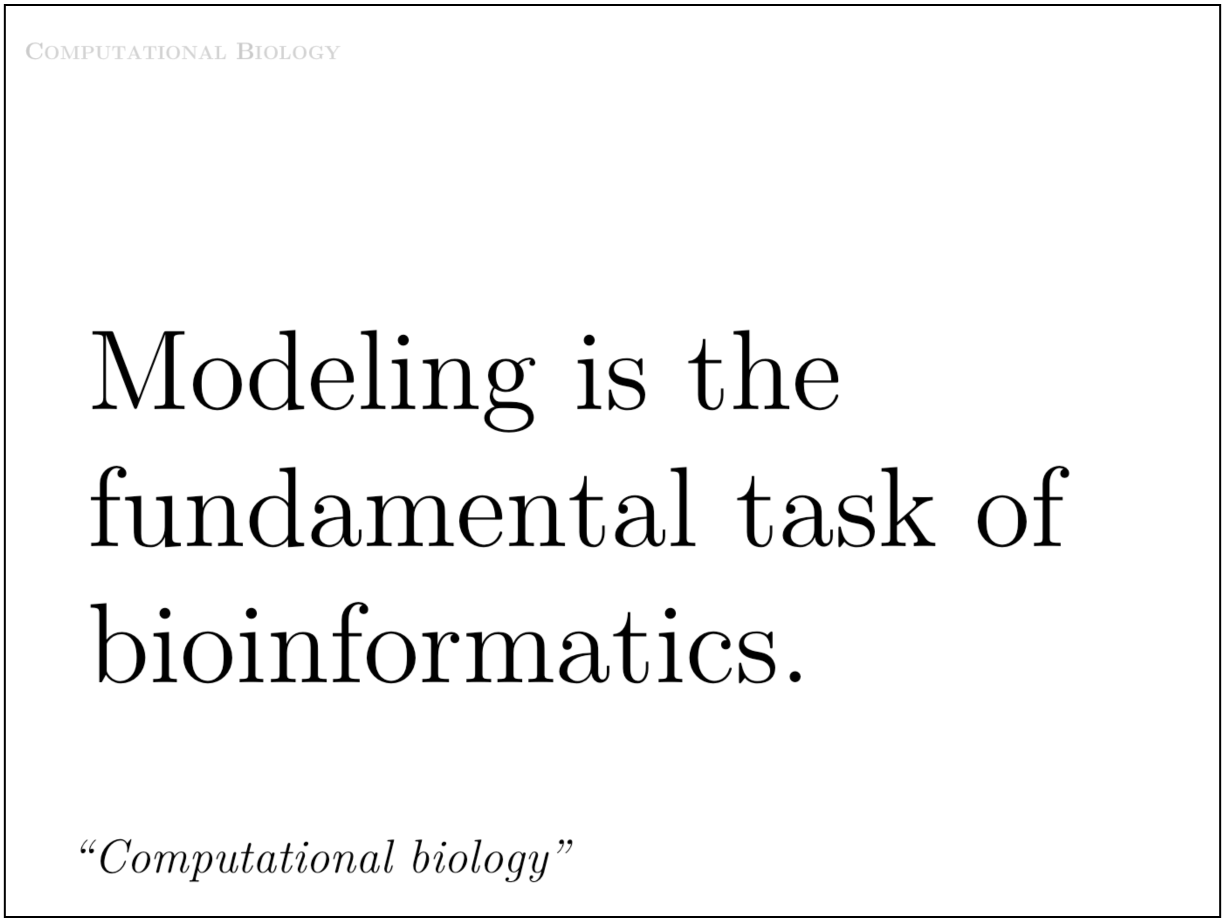
How can bioinformatics help to understand biology?

The question is not: “What can you do?” but: “What should you do?” !

Data alone can however not be translated into progress. **Data does not explain itself.** Modern concepts must inform intelligent analysis; innovative methods must be applied to current data. There is a catch however, in that bioinformatics users should not be required to become bioinformatics experts, rather than "application domain experts". And for those who focus on the applications, i.e. on the biology, it becomes a challenge to keep up with the state-of-the art. What is the best available data repository? What is the best available tool to search for information? What is the best way to access data? It is hard to compare data resources, for example, to rank quality of curation. And by the time we have reviewed all relevant databases, many will already have become obselete.

This problem holds for the data sources, as well as for analysis tools and services.

To address this, we need to focus on objectives, not on methodology.



Computational Biology

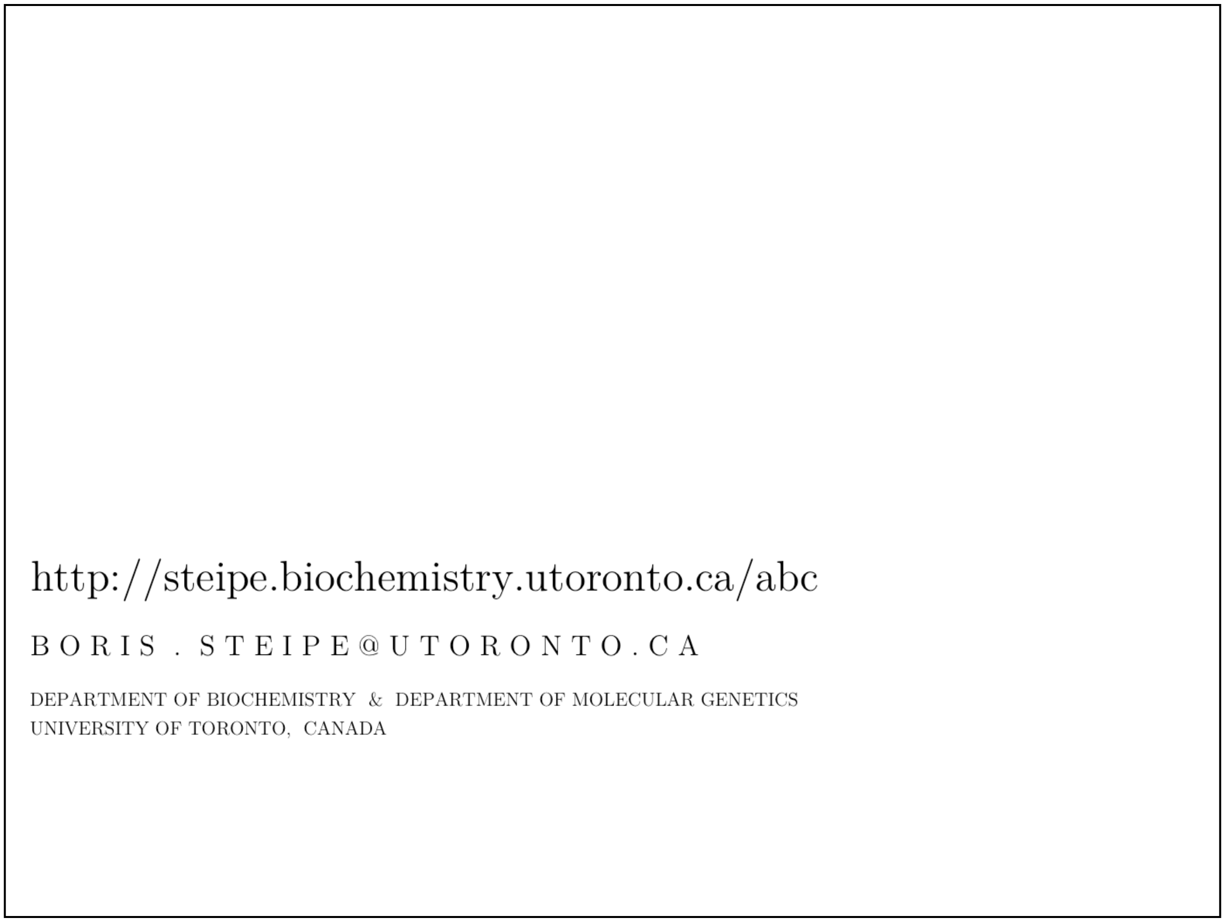
Modeling is the fundamental task of bioinformatics.

“Computational biology”

Looking beyond data management, bioinformatics is a way to study biology. This aspect – which I like to refer to as "Computational Biology" – has a lot more to do with modeling, and with the question of **understanding** biology, than with managing large amounts of data.

Understanding biology means being able to abstract from apparent complexity, in order to interpret observations in the framework of simple, fundamental principles. Such understanding should allow us to make precise, confident predictions.

This involves abstraction, and working with abstractions means we are working with models.



http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA