BINF 630: Bioinformatics Methods

Iosif Vaisman

Email: ivaisman@gmu.edu

Example of a Pubmed entry

```
PMID- 15887224
OWN - NLM
STAT- MEDLINE
DA  - 20050718
DCOM- 20060522
IS  - 1097-0134 (Electronic)
IS  - 0887-3585 (Linking)
DP  - 2005 Aug 15
TI  - New method for protein secondary structure assignment based on a simple
      topological descriptor.
PG  - 513-24
AB  - A simple, five-element descriptor, derived from the Delaunay tessellation of a
      protein structure in a single point per residue representation, can be assigned
      to each residue in the protein. The descriptor characterizes main-chain topology
      and connectivity in the neighborhood of the residue and does not explicitly
      depend on putative hydrogen bonds or any geometric parameter, including bond
      length, angles, and areas. Rules based on this descriptor can be used for
      accurate, robust, and computationally efficient secondary structure assignment
      that correlates well with the existing methods.
AD  - Laboratory for Structural Bioinformatics, School of Computational Sciences,
      George Mason University, Manassas, Virginia 20110, USA.
FAU - Taylor, Todd
AU  - Taylor T
FAU - Rivera, Margarita
AU  - Rivera M
FAU - Wilson, Glenda
AU  - Wilson G
FAU - Vaisman, Iosif I
AU  - Vaisman II
LA  - eng
PT  - Journal Article
PT  - Research Support, U.S. Gov't, Non-P.H.S.
PL  - United States
```

Example of a Pubmed entry

```
TA  - Proteins
JT  - Proteins
JID - 8700181
RN  - 0 (Proteins)
SB  - IM
MH  - Algorithms
MH  - Amino Acid Sequence
MH  - Biophysical Phenomena
MH  - Biophysics
MH  - Computational Biology/*methods
MH  - Computer Simulation
MH  - Hydrogen Bonding
MH  - Models, Chemical
MH  - Models, Molecular
MH  - Molecular Sequence Data
MH  - Protein Conformation
MH  - Protein Folding
MH  - Protein Structure, Secondary
MH  - Protein Structure, Tertiary
MH  - Proteins/chemistry
MH  - Proteomics/*methods
MH  - Sensitivity and Specificity
MH  - Software
MH  - Structural Homology, Protein
EDAT- 2005/05/12 09:00
MHDA- 2006/05/23 09:00
CRDT- 2005/05/12 09:00
```

ID Converters

Features of mapping services	Gene/Clone ID converter	ID mapping by UniProt	MatchMiner	DAVID gene ID conversion tool
Interface	Web-based GUI form	Web-based GUI form	Web-based GUI form, command line	Web-based GUI form
Output format	Html, text, spreadsheet	Html, text	Html, text, spreadsheet	Html, text, spreadsheet
Organisms	Human, mouse, rat	Human, mouse, rat and many other species	Human, mouse	Human, about another 90,000 species
Input/output gene	HUGO gene names, Entrez gene Ids, Ensembl gene Ids, UniGene cluster Ids, RefSeq RNAs (Additional output: CCDS)[a]	Entrez Gene, HGNC, Ensembl, UniGene, TIGR (JCVI)	Gene Symbol HUGO/Alias, Name, UniGene Cluster Id, Entrez Gene Id, RefSeq RNA	Entrez gene Id, Ensembl gene/transcript Id, RefSeq mRNA accession, UniGene Id
Input/output protein	RefSeq peptides, SwissProt names (Additional output: IPI, PDB)*	UniProtKB, RefSeq, GenPept, IPI, PDB	RefSeq protein	PIR accession, PIR Id, PIR NREF Id, RefSeq Protein accession, Uniprot Id/accession, UniRef Id

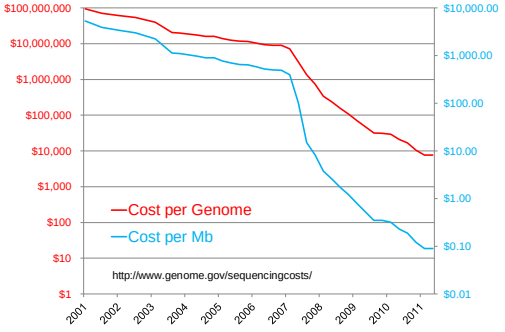
Chavan et al., 2011

ID Converters

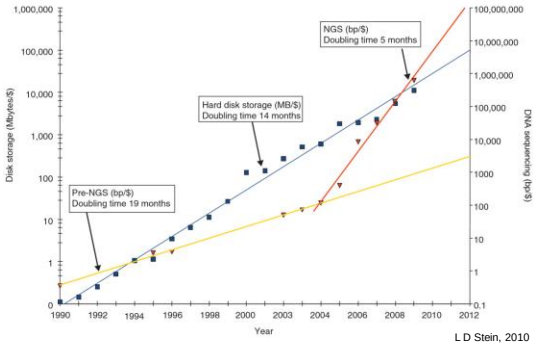
Mapping services	Link
Gene/Clone ID converter	http://idconverter.bioinfo.cnio.es/
ID mapping by UniProt	http://www.uniprot.org/?tab=mapping
MatchMiner	http://discover.nci.nih.gov/matchminer/
DAVID gene ID conversion tool	http://david.abcc.ncifcrf.gov/conversion.jsp
g:Convert	http://biit.cs.ut.ee/gprofiler/gconvert.cgi
CRONOS	http://mips.helmholtz-muenchen.de/genre/proj/cronos/
bioDBnet:db2db	http://biodbnet.abcc.ncifcrf.gov/db/db2db.php

Chavan et al., 2011

Genome sequencing costs

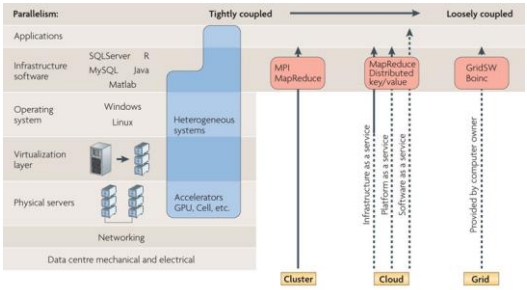


Sequencing and storage cost



L D Stein, 2010

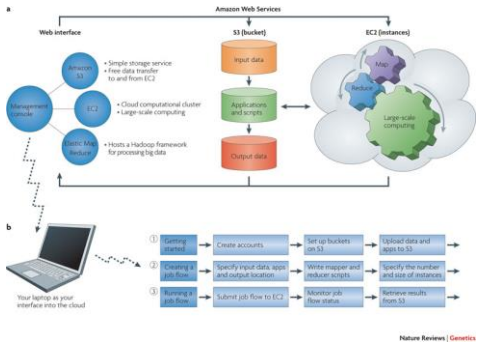
Infrastructure organization



Nature Reviews | Genetics

Schadt et al., 2010

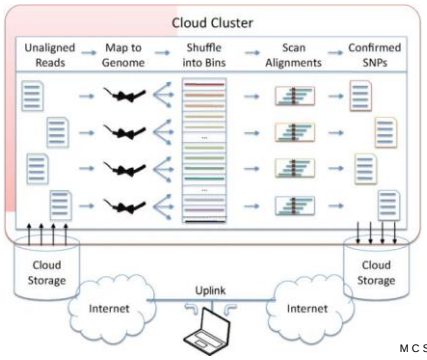
Cloud computing



Nature Reviews | Genetics

Schadt et al., 2010

Cloud computing and DNA sequencing



M C Schatz et al., 2010

Types of computational environments

Large-scale computing platform	Computing architectures	Advantages	Disadvantages	Example applications
Cluster computing	Multiple computers linked together, typically through a fast local area network, that effectively function as a single computer	Cost-effective way to realize supercomputer performance	Requires a dedicated, specialized facility, hardware, system administrators and IT support	• BLAST • Bayesian network reconstruction • Computing genetic associations in large-scale GWA studies
Cloud computing	Computing capability that abstracts the underlying hardware architectures (for example, servers, storage and networking), enabling convenient, on-demand network access to a shared pool of computing resources that can be readily provisioned and released (NIST, technical Report)	The virtualization technology used results in extreme flexibility; good for one-off HPC tasks, for which persistent resources are not necessary	Privacy concerns; less control over processes; bandwidth is limited as large data sets need to be moved to the cloud before processing	• Searching sequence databases • Aligning raw sequencing reads to genomes • General purpose genomics tools (for example, GeneSifter from Geospiza) • Most applications running on a cluster can be transferred to a cloud
Grid computing	A combination of loosely coupled networked computers from different administrative centres that work together on common computational tasks. Typified by volunteer computing efforts (such as Folding@Home), which 'scavenge' spare computational cycles from volunteers' computers	Ability to enlist large-scale computational resources at low or no cost (large-scale volunteer-based efforts)	Big data transfers are difficult or impossible; minimal control over underlying hardware, including availability	• Protein folding (Folding@Home) • Proteome analysis • Protein prediction (Rosetta@Home) • Predicting interactions between small molecules and proteins (FightAIDS@Home) • Condor project
Heterogeneous computing	Computers that integrate specialized accelerators — for example, GPUs or reconfigurable logic (FPGAs) — alongside GPPs	Cluster-scale computing for a fraction of the cost of a cluster; optimized for computationally intensive fine-grained parallelism; local control of data and processes	Significant expertise and programmer time required to implement applications; not generally available in cluster- and cloud-based services	• Bayesian network learning • Protein folding (Folding@Home) • Molecular dynamics simulation (NAMD) • BLAST • CLUSTALW • HMMER • Reconstruction of evolutionary trees

The above categories are not exclusive. For example, heterogeneous computers are often used as the building blocks of cluster, grid or cloud computing systems; the shared computational clusters available in many organizations could be described as private Platform as a Service (PaaS) clouds. The main differences between the platforms are degree of coupling and tenancy — grid and cloud computers are designed for loosely coupled parallel workloads, with the grid resources allocated exclusively for a single user, whereas the underlying hardware resources in the cloud are typically shared among many users (multi-tenancy). Cluster computers are typically used for tightly coupled workloads and are often allocated to a single user. FPGA, field-programmable gate array; GPP, general purpose processor; GPU, graphics processing unit; GWA, genome-wide association; HPC, high-performance computing; NIST, National Institute of Standards and Technology.

Schadt et al., 2010

Types of computational environments

Environment	URL
Cloud computing	
Amazon Elastic Compute Cloud	http://aws.amazon.com/ec2
Bionimbus	http://www.bionimbus.org
NSF CluE	http://www.nsf.gov/cise/clue/index.jsp
Rackspace	http://www.rackspacecloud.com
Science Clouds	http://www.scienceclouds.org
Heterogeneous computing	
NVIDIA GPUs	http://www.nvidia.com
AMD/ATI GPUs	http://www.amd.com
Heterogeneous cloud computing	
SGI Cyclone Cloud	http://www.sgi.com/products/hpc_cloud/cyclone
Penguin Computing On Demand	http://www.penguincomputing.com/POD/Summary

GPU, graphics processing unit; NSF, US National Science Foundation.

Schadt et al., 2010

Defining Big Data

NOT JUST SIZE

The three Vs of Big Data: volume, variety and velocity
(D.Laney, 2001)

Elements of "Big Data" include:

- The degree of complexity within the data set
- The amount of value that can be derived from innovative vs. non-innovative analysis techniques
- The use of longitudinal information supplements the analysis

http://mike2.openmethodology.org/wiki/Big_Data_Definition

Data Mining

- Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules
- Common data mining tasks
  - Classification
  - Estimation
  - Prediction
  - Affinity Grouping
  - Clustering
  - Description

Knowledge Discovery

Knowledge is a.pattern that exceeds certain threshold of interestingness.

Factors that contribute to interestingness:
coverage
confidence
statistical significance
simplicity
unexpectedness
actionability

Knowledge Discovery

- Directed and Undirected KD
- Directed KD
  - Purpose: Explain value of some field in terms of all the others
  - Method: We select the target field based on some hypothesis about the data. We ask the algorithm to tell us how to predict or classify it
  - Similar to hypothesis testing (e.g., in regression modeling) in statistics

Knowledge Discovery

- Undirected KD
  - Purpose: Find patterns in the data that may be interesting
  - Method: clustering, affinity grouping
  - Closest to ideas of machine learning in artificial intelligence
- Comparison
  - UKD helps us to recognize relationships & DKD helps us to explain them

Classification

- Classifying observations into different categories given characteristics

Estimation

- Rules that explain how to estimate a value given characteristics

Prediction

- Rules that explain how to predict a future value or classification, given characteristics

Affinity Grouping

- Grouping by relations (not by characteristics)

Clustering

- Segmenting a diverse population into more similar groups
- In clustering, there are no pre-defined classes and no examples. Records are grouped together by some similarity measure.

Knowledge Discovery