Bioinformatics Topics

Data Generation

Experimental Data types include :

Sequences - Typically Next-Generation DNA Sequencing (NGS).

3D Protein Structures - X-ray crystallography or Nuclear magnetic resonance spectroscopy (NMR)

Gene Expression Data - Microarrays

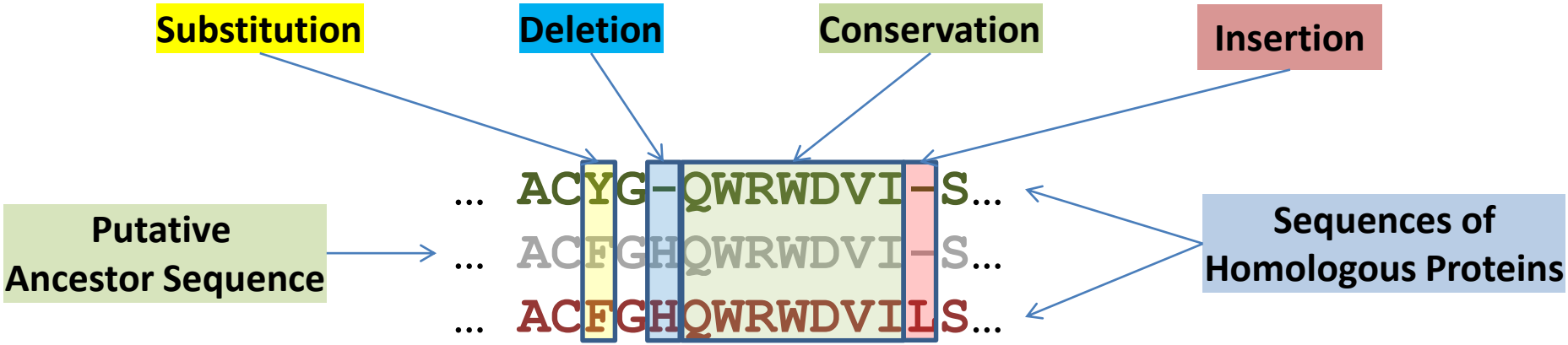
Bioinformatics Topics

Data Generation

Data Analysis

The Alignment of Pairs of Homologous DNA/Protein sequences.

Fundamental to most forms of DNA/Protein Sequence analysis



Bioinformatics Topics

Data Generation

Data Analysis

The Alignment of Families of Homologous sequences.

First, find a family of Homologous sequences.

```
... APFELVISWKLIVESPAINCDWRTENGLANDSGMLVNOWPAI ...
... APYELVISQWKLIVESNPAINKDWRTYENGLANDSGMLVNOWAI ...
... APFELVISWKLIVESNPAINCDWRTENGLANDSGMLVNOWAI ...
... APFELVISQWKLIVESNPAINCDWRTENGLANDSGMLVNOWAI ...
... APYELVISWKLIVESNPINCDWRTENGLANDRSGMLINOWAI ...
... APFELVISQWKLIVESNPAINCDWRTENGLANDSGMLVNOWLI ...
... APFELVISQWKLIVESNPAINDWRTENGLANDSGMLVNOWAI ...
... APYELVISWKLIVESNPAINCDWRTENGLANDSGMLLNOWMI ...
```

Bioinformatics Topics

Data Generation

Data Analysis

The Alignment of Families of Homologous sequences.

Then, align by inserting "–"s representing InDels, in each sequence.

```
... APFELVIS-WKLIVES-PAINCDWRT-ENGLANDSGMLV-NOWPAI ...
... APYELVISQWKLIVESNPAINKDWRTYENGLANDSGMLV-NOW-AI ...
... APFELVIS-WRLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNP-INCDWRT-ENGLANDRSGMLINOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-LI ...
... APFELVISQWKLIVESNPAIN-DWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNPAINCDWRT-ENGLANDSGMLL-NOW-MI ...
```

Bioinformatics Topics

Data Generation

Data Analysis

The Alignment of Families of Homologous sequences.

Next, identify the columns where Substitutions and/or InDels have been predicted.

```
... APFELVIS-WKLIVES-PAINCDWRT-ENGLANDSGMLV-NOWPAI ...
... APYELVISQWKLIVESNPAINKDWRTYENGLANDSGMLV-NOW-AI ...
... APFELVIS-WRLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNP-INCDWRT-ENGLANDRSGMLINOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-LI ...
... APFELVISQWKLIVESNPAIN-DWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNPAINCDWRT-ENGLANDSGMLL-NOW-MI ...
```

Bioinformatics Topics

Data Generation

Data Analysis

The Alignment of Families of Homologous sequences.

Then, identify the columns where full Conservation has been predicted.

```
... APFELVIS-WKLIVES-PAINCDWRT-ENGLANDSGMLV-NOWPAI ...
... APYELVISQWKLIVESNPAINKDWRTYENGLANDSGMLV-NOW-AI ...
... APFELVIS-WRLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNP-INCDWRT-ENGLANDRSGMLINOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-LI ...
... APFELVISQWKLIVESNPAIN-DWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNPAINCDWRT-ENGLANDSGMLL-NOW-MI ...
```

Bioinformatics Topics

Data Generation

Data Analysis

The Alignment of Families of Homologous sequences.

Finally … Identify the Glorious Message!!!!.

```
... APFELVIS-WKLIVES-PAINCDWRT-ENGLANDSGMLV-NOWPAI ...
... APYELVISQWKLIVESNPAINKDWRTYENGLANDSGMLV-NOW-AI ...
... APFELVIS-WRLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNP-INCDWRT-ENGLANDRSGMLINOW-AI ...
... APFELVISQWKLIVESNPAINCDWRT-ENGLANDSGMLV-NOW-LI ...
... APFELVISQWKLIVESNPAIN-DWRT-ENGLANDSGMLV-NOW-AI ...
... APYELVIS-WKLIVESNPAINCDWRT-ENGLANDSGMLL-NOW-MI ...
```

Bioinformatics Topics

Data Generation

Data Analysis

Searching for Homologous Sequences in a Sequence Database.

Database searching is the most common Bioinformatics process by far.

Database searching is pairwise comparison repeated many times.

Non-optimal comparison methods are essential for practical reasons.

A list of matches, ordered by the improbability of occurring just by chance is generated.

Bioinformatics Topics

Data Generation

Data Analysis

Searching for Homologous Sequences in a Sequence Database.

Database searching seeks “Similarity”. Users seek “Homology”.

```
Query            KLYPLRPQTPEPPPPPPPPPPPPLPAAPPQP
Similarity       +L P +P     P P PP P PP      PP+P
Database Entry   RLTPPQPLMMPPRPTPPTPLPPATLTVPPRP
```

Homology?

Or 2 proteins including a lot of Prolines??

Bioinformatics Topics

Data Generation

Data Analysis

Searching for Homologous Sequences in a Sequence Database.

Database searching seeks “Similarity”. Users seek “Homology”.

```
Query            TCTCCCATTCGTAAAAAAAAAAAAAAAAA
Similarity       ||| ||||| ||||||||||||||||||||
Database Entry   TCTTCCATTTGTAAAAAAAAAAAAAAAAA
```

Homology?

Or 2 unrelated mRNAs??

Bioinformatics Topics

Data Generation

Data Analysis

Searching for Homologous Sequences in a Sequence Database.

Database searching seeks “Similarity”. Users seek “Homology”.

```
Query            TTAGCAAGATCAGCCCTAACTCGGCATCTT
Similarity         |  ||   |  ||  |  |  |  ||   |  ||   |
Database Entry   CTTGCGCGCTCTGTCTTGACGAGACACTTA
```

Homology?

A very unconvincing alignment!!

Bioinformatics Topics

Data Generation

Data Analysis

Searching for Homologous Sequences in a Sequence Database.

Database searching seeks “Similarity”. Users seek “Homology”.

TTAGCAAGATCAGCCCTAACTCGGCATCTT

Query	L	A	R	S	C	L	T	R	H	L
Similarity	L	A	R	S	C	L	T	R	H	L
Database Entry	L	A	R	S	C	L	T	R	H	L

CTTGCGCGCTCTGTCTTGACGAGACACTTA

Homology?

Probable --- a perfect protein match??

In all circumstances – always align at the protein level wherever possible.

Bioinformatics Topics

Data Generation

Data Analysis

Searching for simple sequence patterns Sequences in DNA.

Largely a matter of finding short sequences within longer ones.

Computationally trivial.

A concrete example is required:

Restriction Mapping

Detecting Restriction Enzyme Recognition Sites is complicated by their redundancy.

Few Recognition Sites can be simply defined using only the codes A, C, G and T.

The solution is to use the Nucleotide Ambiguity Codes defined by IUPAC.

Unambiguous site (EcoRI): G/AATC

Ambiguous site (PpuMI): RG/GWCCY

Cut here

And here

TTAGCAAGATCAGGACCTACTCGGCATCTTCCTGGGTCCC

RGGWCCY

IUPAC DNA Alphabet

Code	Meaning
A	A
C	C
G	G
T/U	T/U
M `aMino`	A|C
R `puRine`	A|G
W `Weak`	A|T
S `Strong`	C|G
Y `pYrimidine`	C|T
K `Keto`	G|T
V `not T`	A|C|G
H `not G`	A|C|T
D `not C`	A|G|T
B `not A`	C|G|T
N `aNy`	A|C|G|T

Bioinformatics Topics

Data Generation

Data Analysis

Searching for simple sequence patterns Sequences in Proteins.

Patterns can be derived manually to represent conserved regions of MSAs

Simple where conservation is 100%

```
... CQVLNPYYHWGQCGGIGWSGPTVCASGTT ...
... CQYSNDYYHWGQCGGIGWSGCKTCTSGTT ...
... CHVLNPYYQWGQCGGIGWTPSTTCASPYT ...
... CSTLNPYYVWGQCGGIGWSGPTNCAPGSA ...
... CVYSNDYYVWGQCGGIGWSGPTCCASGST ...
```

WGQCGGIGW ← Pattern

Bioinformatics Topics

Data Generation

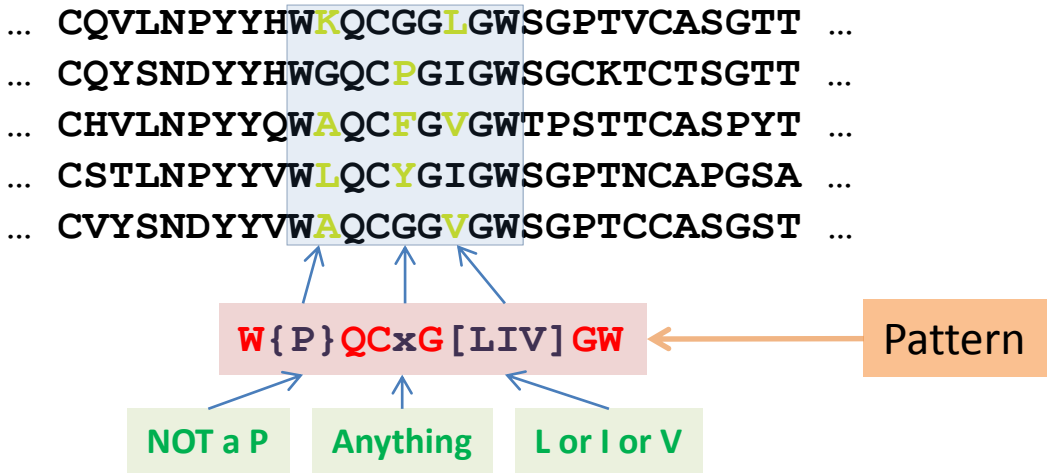
Data Analysis

Searching for simple sequence patterns Sequences in Proteins.

Not so easy where conservation is less than perfect

An Amino Acid Alphabet including all ambiguities is not practical!

The solution is a simple syntax for ambiguous amino acid sequences.



Bioinformatics Topics

Data Generation

Data Analysis

Searching for simple sequence patterns Sequences in Proteins.

Bioinformatics Topics

Data Generation

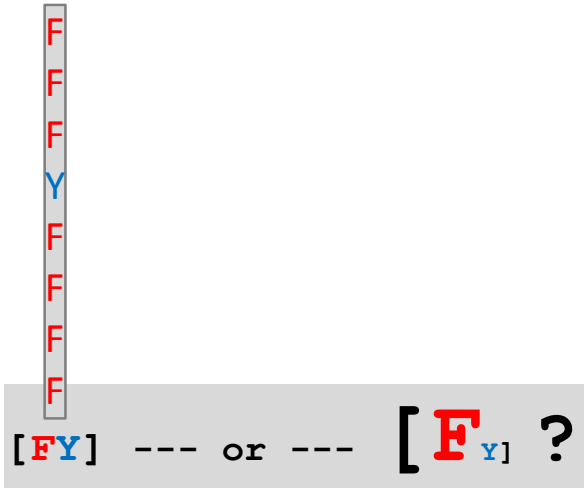
Data Analysis

Searching for simple sequence patterns Sequences in Proteins.

Simple Protein patterns are of limited precision.

Only highly conserved regions can be described usefully.

Patterns cannot weight possibilities by frequency.



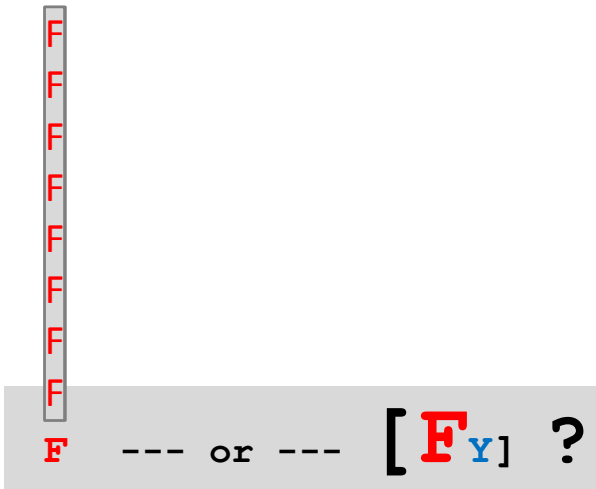
Bioinformatics Topics

Data Generation

Data Analysis

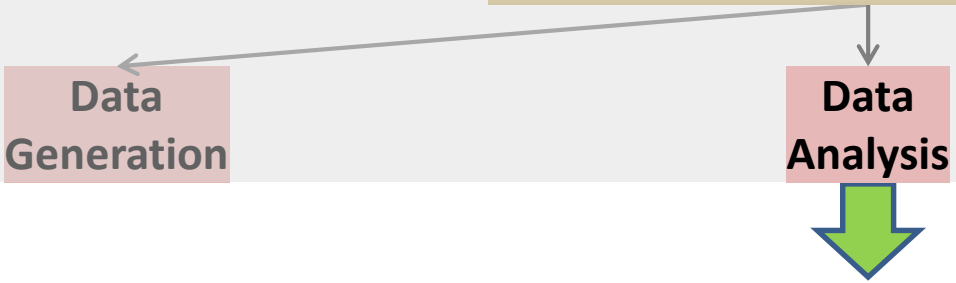
Searching for simple sequence patterns Sequences in Proteins.

Simple Protein patterns are of limited precision.



Patterns do not reflect commonly accepted substitutions.

Bioinformatics Topics



Searching for Protein properties with better models.

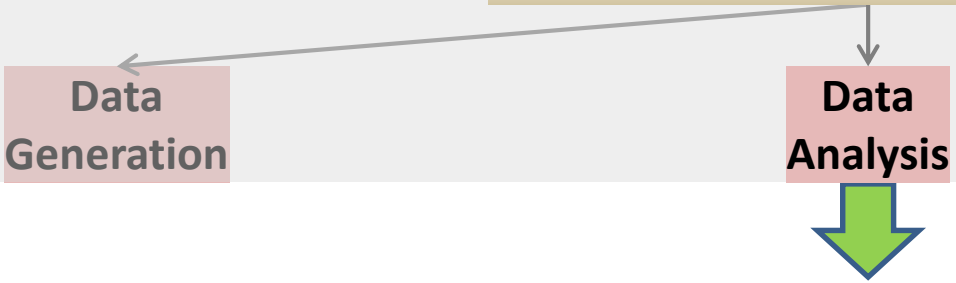
Again, start with an MSA of instances of the feature to be modelled.

Create a “suitable” representation of the relevant portion of MSA

Compare the model along other protein sequences was illustrated for simple patterns.

Where matches are detected, the corresponding protein property is likely to occur.

Bioinformatics Topics



Searching for Protein properties with better models.

A variety of simple models have been developed (e.g. Position Weight Matrices) for a number of purposes, including:

- Gene discovery in bacteria genomes (DNA)
- Early versions of 2D protein Structure Prediction
- Transmembrane Alpha Helix prediction
- TATA box Detection (DNA)
- Helix-Turn-Helix (HTH) Prediction
- Prediction of Coiled Coils

The most powerful and prolific current profiles are Hidden Markov Models (HMMs)

Bioinformatics Topics

Data Generation

Data Analysis

Estimating evolution - Phylogeny.

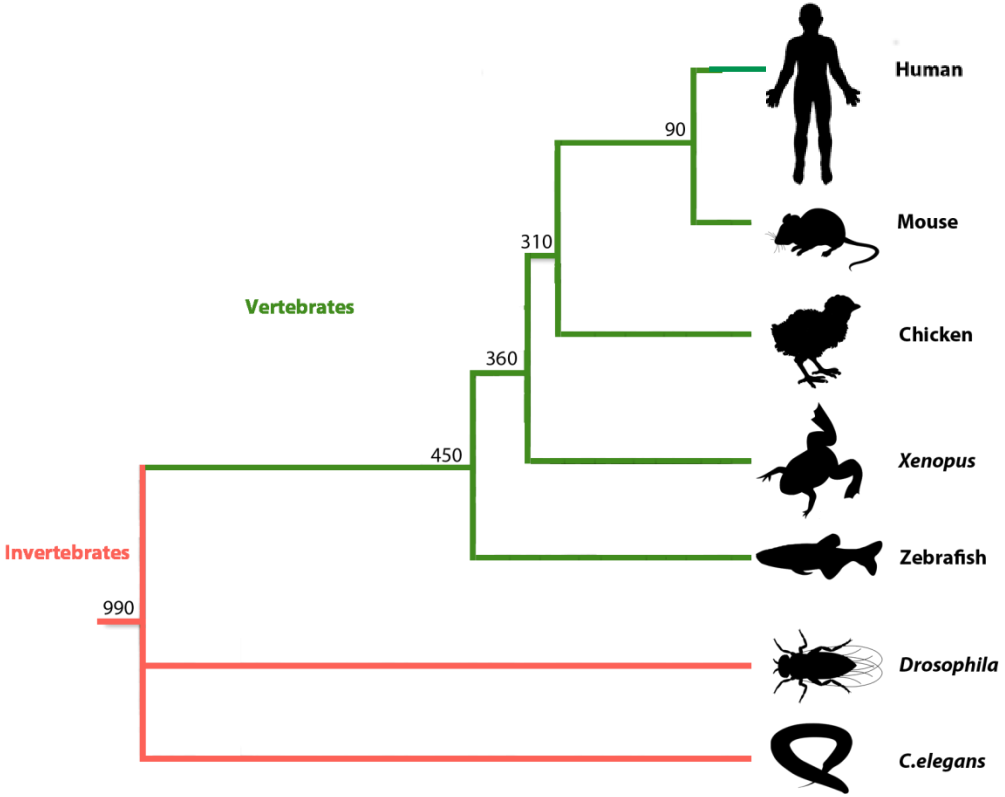
Broadly, the estimation of evolutionary history from available evidence.

"Evidence" does not *have* to be a carefully crafted MSA of Orthologous sequences from a range of organisms.

However, in the context of Bioinformatics, it invariably is.

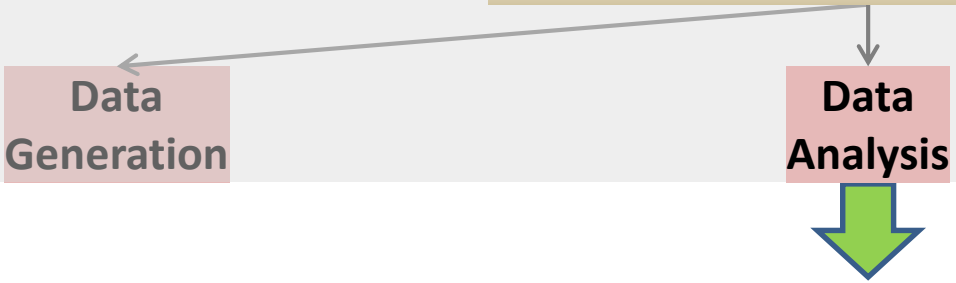
Typically, conclusions of Phylogenetic analysis are represented as Evolutionary Trees.

Which are very Beautiful!!



My personal preference is for trees that place ***ME*** as far away from a ***MOUSE*** as possible!!!!

Bioinformatics Topics



Estimating evolution - Phylogeny.

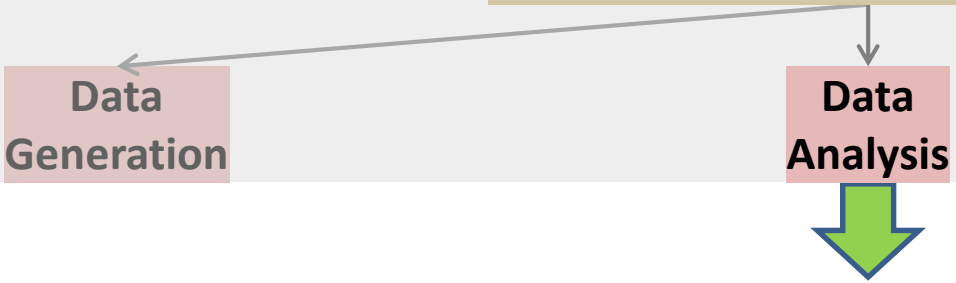
Phylogeny is another example of an analysis based on MSAs.

One very effective Phylogenetic strategy is to seek an answer to the question:

"What is the most probable Evolutionary Tree, given I believe this MSA to be perfect?"

Reinforcing how central is the role of Statistics in Bioinformatics.

Bioinformatics Topics



Protein structure prediction. Secondary Structure.

Essentially predicting the locations of Alpha Helices, Beta Sheets and Turns.

Modern methods employ Machine Learning to generate Artificial Neural Networks.

That is profiles computed by “learning” from observation of examples.

Bioinformatics Topics

Data Generation

Data Analysis

Protein structure prediction. Secondary Structure.

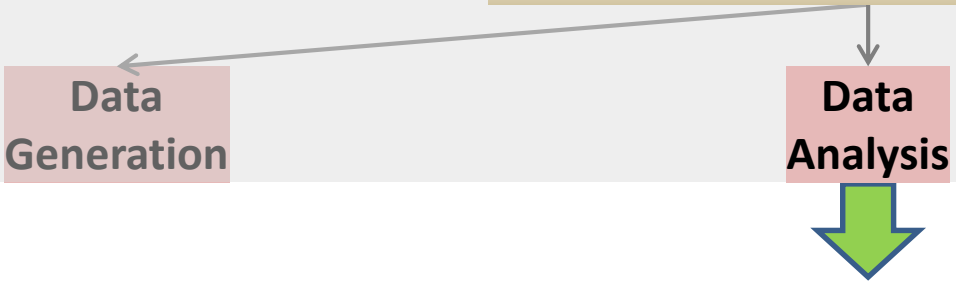
Better predictions are obtained from MSA data than from individual protein sequences.

General principle being, the more information offered, the more reliable the prediction.

Some systems will automatically generate an MSA if offered a solitary protein sequence.

Prediction will be based on the MSA, computed by iterative database searching.

Bioinformatics Topics



Protein structure prediction. Tertiary Structure.

Predicting Tertiary Structure directly from Primary Structure is not currently practical.

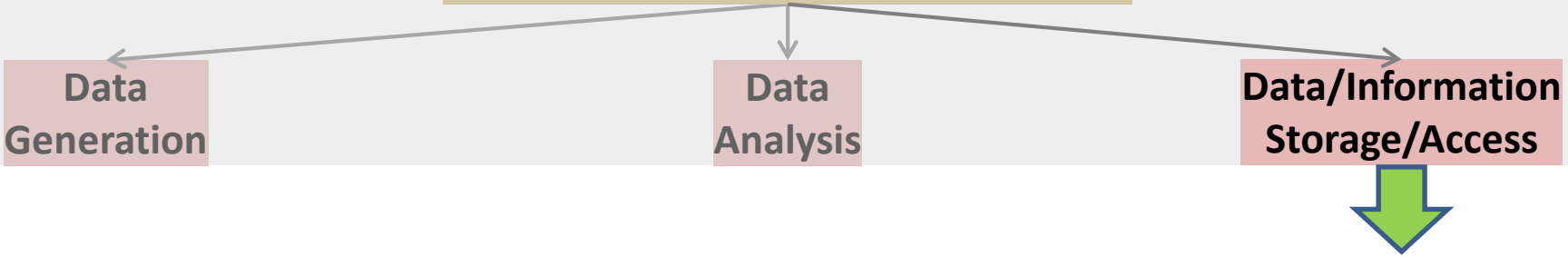
De novo protein structure prediction requires better algorithms and more computing power.

Homology modelling requires a reliable Tertiary Structure for a homologous protein.

Tertiary Structure for a protein is predicted by comparison with the homologous structure.

Homology modelling is hampered by low volumes and uneven spread of available structures.

Bioinformatics Topics



Overview.

Raw Experimental Data, can next be Annotated in the light of analytical revelation.

Data + Annotation = Information.

Information can now be stored in Databases that allow users easy and *unrestricted* access.

Primary DNA Sequence Databases

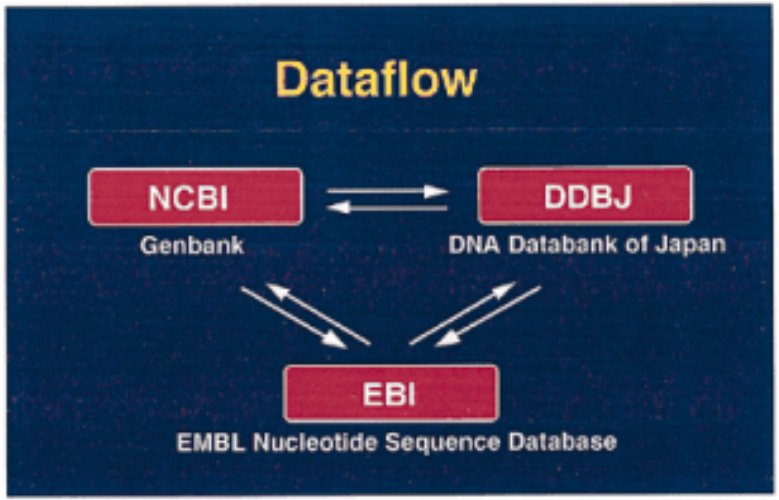
Original submission by experimentalists
Content controlled by the submitter

EMBL NUCLEOTIDE SEQUENCE DATABASE

NCBI GenBank

DDBJ DNA Data Bank of Japan

INSDC International Nucleotide Sequence Database Collaboration



Primary Protein Sequence Databases

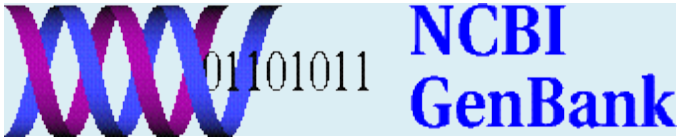
UniProtKB

an encyclopedia on proteins

composed of 2 sections
UniProtKB/TrEMBL and UniProtKB/Swiss-Prot
unreviewed and reviewed
automatically annotated and manually annotated

Derivative Sequence Databases

Built from primary data



RefSeq

Submission by experimentalists
Significant redundancy
Annotation inconsistent
DNA and RNA only

non-redundant
richly annotated
DNA, RNA, protein
diverse taxa

akin to the primary research literature

akin to the review literature

Derivative Databases for Protein Features

Collections of HMMs representing Protein Domains and/or Motifs derived from Protein sequence Databases.

Derivative Databases for Protein Features

It is generally wise to use more than one Feature Searching service.

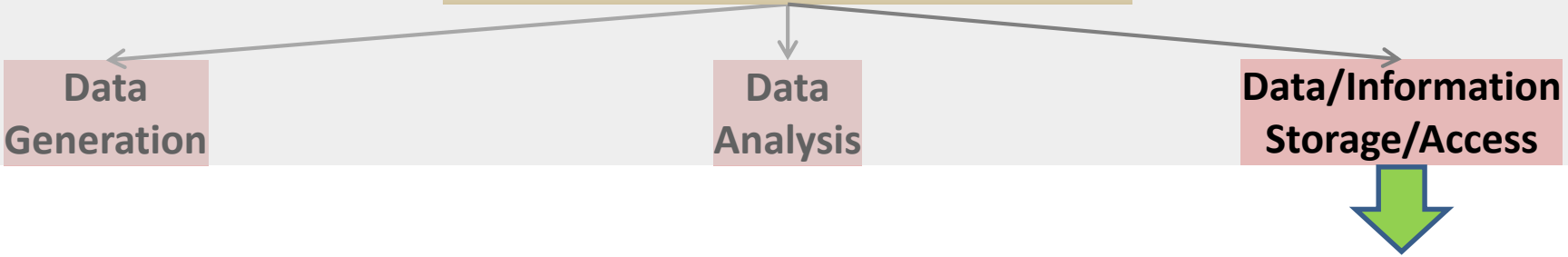
This can be tedious, involving many websites and different search tools.

is a consortium of member databases.

defines protein families, domains, regions, repeats and sites according to matches against member databases

enables any subset of member databases to be searched together

Bioinformatics Topics



Genome Databases.

Genome Databases store entire genome sequence(s) AND their interpretation.

Each new sequenced genome or significantly re-assembled existing genome is fully analysed.

The individual processes for manual analysis are the same as those for automatic analysis.
Most have been mentioned in this simple talk.

Analysing an individual gene can be done manually.
Analysing an entire genome is only practical using automated strategies.

Bioinformatics Topics

Data Generation

Data Analysis

Data/Information Storage/Access

Genome Databases.

The Three foremost Genome Database options

Ensembl and UCSC Browser software can be downloaded and used for private datasets.

Bioinformatics Topics

- Data Generation
- Data Analysis
- Data/Information Storage/Access

Protein Structure Databases.

Worldwide Protein Data Bank Foundation

RCSB PDB PROTEIN DATA BANK

PDBj Protein Data Bank Japan

PDBe Protein Data Bank in Europe

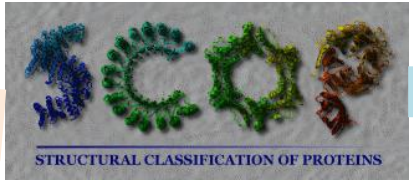
Bioinformatics Topics

Data Generation

Data Analysis

Data/Information Storage/Access

Protein Structure Databases.



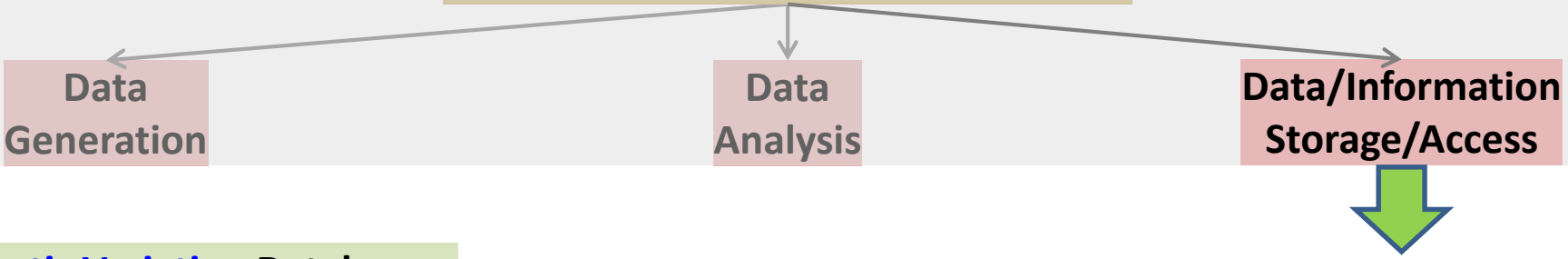
Superfamily

HMM library and genome assignments server

Worldwide Protein Data Bank Foundation

CATH Gene3D

Bioinformatics Topics



Genetic Variation Databases.

Databases storing the many genetic variations that occur between individuals and species.

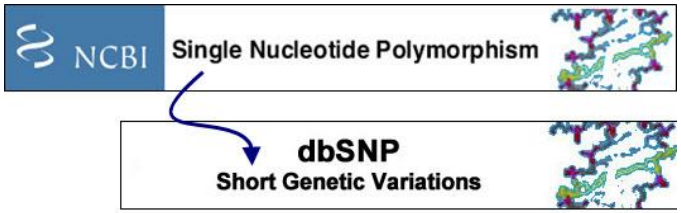
Widely incorporated into Genome Databases, such as Ensembl.

Since High Throughput Sequencing (HTS) has become standard, variation detection has become easier. Databases have developed dramatically.

Bioinformatics Topics

Data Generation | Data Analysis | **Data/Information Storage/Access**

Genetic Variation Databases.



dbSNP is the largest general database for genetic variations.

Originally just Single Nucleotide Polymophisms (SNPs).

Now includes other types of Short Genetic Variation.

dbSNP, originally focused on human variations, now covers many organisms.

dbSNP now records relationships between variation and phenotype.

Bioinformatics Topics

Data Generation

Data Analysis

Data/Information Storage/Access

Other relevant databases include:

Microarray databases

There are a considerable number, both commercial and public domain.

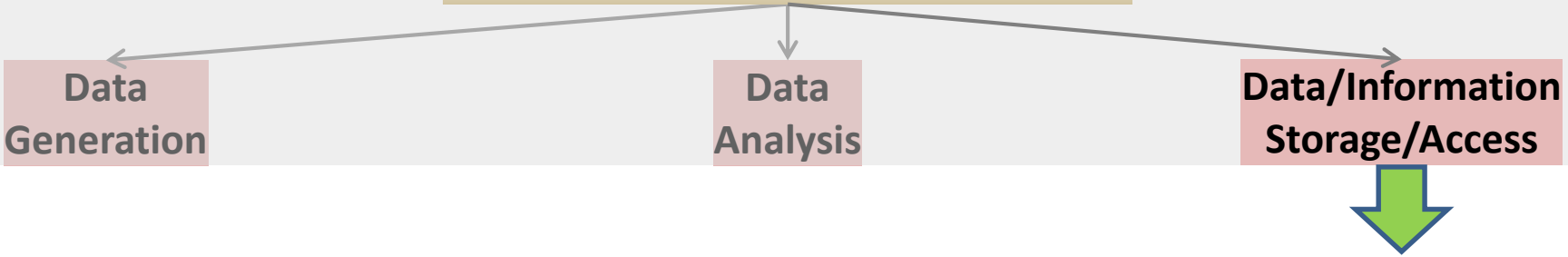
Two major Public Domain Microarray Databases are:

The Gene Expression Omnibus (GEO), maintained in America.

Gene Expression Omnibus

ArrayExpress, maintained in Europe.

Bioinformatics Topics



Other relevant databases include:

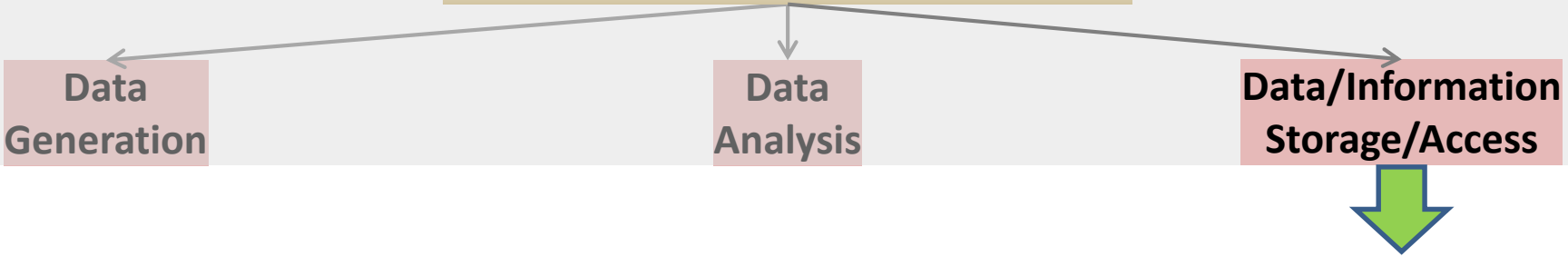
Microarray databases

High Throughput Sequencing (HTS) has become a viable option to the use of Microarrays.

Accordingly, both GEO and ArrayExpress now manage HTS data sets.

ArrayExpress regularly imports data from GEO.

Bioinformatics Topics



Other relevant databases include:

Literature databases

Many free literature search/access services are available via the INTERNET.

You will be introduced to, arguably, the best and most famous as a part of this course.

Bioinformatics Topics

Data Generation

Data Analysis

Data/Information Storage/Access

Other relevant databases include:

[Gene Ontology Database](#)

Early Primary Database annotation was poor.

Annotation was left to the submitted and then not curated .

In consequence, Database Searching just by Keyword was far from reliable.

Bioinformatics Topics

Data Generation

Data Analysis

Data/Information Storage/Access

Other relevant databases include:

Gene Ontology Database

The Gene Ontology (GO) database provides a hierarchy of formally agreed terms to describe gene products accurately and unambiguously.

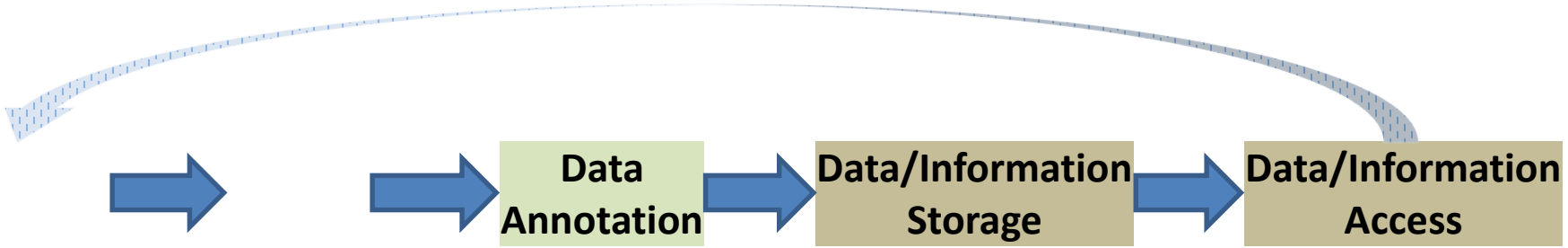
Searching with these terms radically improves the efficacy of annotation searching.

Bioinformatics Topics

Data Generation

Data Analysis

Data/Information Storage/Access



A simplistic ordering for the Bioinformatics Topics discussed here

And now … Once again … Your turn!
Some issue for consideration, discussion and reaction

Define the three terms Homologue, Paralogue and Orthologue, being ever assiduous to ignore offensive American misspellings!

The is but one basic strategy for computing Pairwise Alignments that is considered optimal. However, this strategy can be implemented to compute either Global Alignments or Local Alignments.
Just informally, how do these two possibilities differ?

Generally speaking, would you compute MSAs using a Global or a Local approach? Briefly justify your choice.

Generally speaking, would you conduct Database Similarity searches using a Global or a Local approach? Briefly justify your choice.

"Sequence alignment only makes sense for sequences representing Homologous entities"

A profound observation made by the ever sagacious David Philip Judge whilst sipping an eventide cup of Tesco's very cheapest tea in the penthouse suite of his Ivory Tower (personal communication, 2016.06.10).

Consider and comment upon this fundamental truth.

"A Multiple Alignment of Homologous sequences which were a mixture of Orthologues and Paralogues would not be suitable as input data for Phylogenetic analysis "

Another deep one from DPJ

Consider and comment upon this further pearl of enlightenment.

In the course of the dialogue for this presentation, there was mention of "Accepted Substitutions", more formally referred to as "Accepted Point Mutations", or … if you enjoy clumsy for the sake of a pronounceable acronym, "Point Accepted Mutation" (PAM).

How would you informally define an "Accepted Point Mutation"?

The Extended syntax for ScanProsite is the most common syntax used for protein pattern definition. ScanProsite being the program for searching the of the Prosite database. Prosite was first created way back in the 1980s and, initially, was composed exclusively of protein patterns.

There is no great value, at this stage, to be entirely familiar with this very simple syntax. However, from the hints in this presentation and a quick glance at the appropriate web pages, can you interpret the pattern?

```
C{P}x(3,7)[FY](2)Wx(2)[VIL]
```

In the slides preceding, Protein Domains and Protein Sequence Motifs were mentioned with rather sparse explanation.

Define both of these terms and describe simply the difference between them.

In the slide notes, there is mention of Position Weight Matrices (PWMs).

Can you say, simply, what a Position Weight Matrix might be and how it might be used?

What obvious property does a PWM possess that is lacking in a simple sequence pattern (or consensus sequence)?

The best secondary structure programs are reckoned to be around 80% accurate.

It is further suggested that 80% is about as good as it is possible to achieve.

Stated simply, why would you suppose that 100% accuracy might be unobtainable?

Hint: Do you think that two human experts, given the very best evidence of Tertiary Structure, would also agree upon the exact amino acid positions where an Alpha Helix starts and finishes?

Homology Modelling is mentioned in the slides as a method for predicting tertiary structure when structure(s) of protein(s) homologous to the query protein are available. The process involves aligning the query protein with the known structure, using the known sequence as a guide.

It is also possible to predict Tertiary Structure when, known structures thought to be appropriate exist, but only for sequences that ARE NOT HOMOLOGOUS. In such cases, the Primary Sequence corresponding to the known structure will be of little assistance.

Tricky eh!? What are the name(s) for those types of method? ONLY if you can do so VERY simply. Say a few words to say how they over come the lack of a homologous sequence.

It was noted in the slides that often different Protein Feature searches often do not exactly agree.

It is common for two services to agree upon the presence of a domain, but not upon it precise start and end positions within a protein.

Would you find this to be worrying? Surprising? If not, why not?

THE END

BREAK!

More to come I fear … but time for a swift cup of tea perchance?

Maybe time for a short jig? The whistling of a merry tune?

Or, mayhap, a delving into the melodic possibilities of youtube?
There be much good stuff there … I offer you a few of my favourites.

Once fully refreshed …. Click on mon braves!