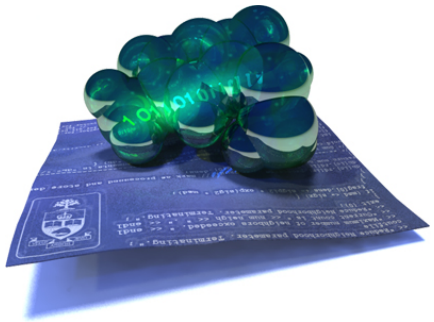
A
Bioinformatics
Course

Concepts of Sequence Analysis



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

What is a sequence? ... or rather

What's *in* a sequence?

```
>unknown sequence
wasisteinnamewasunsroseheisstwieesauchhi
essewuerdelieblichduften
```

What's *in* a sequence?

```
>unknown sequence
wasisteinnamewasunsroseheisstwieesauchhi
essewuerdelieblichduften
```

Analysis by composition:	Analysis by comparison:	Analysis of conservation:
Count components, sum properties.	Compare patterns with an annotated dictionary.	Identify relatedness, analyze change.

What's *in* a sequence?

```
>unknown sequence
wasisteinnamewasunsroseheisstwieesauchhi
essewuerdelieblichduften
```

Analysis by composition:	Analysis by comparison:	Analysis of conservation:
Count components, sum properties.	Compare patterns with an annotated dictionary.	Identify relatedness, analyze change.

```
w-as ist ein - name
| |:  |   ||   ||||
what-'s- -in a name
```

```
>unknown sequence
wasisteinnamewasunsroseheisstwieesauchhi
essewuerdelieblichduften
```

Analysis by composition:	Analysis by comparison:	Analysis of conservation:
Count components, sum properties.	Compare patterns with an annotated dictionary.	Identify relatedness, analyze change.

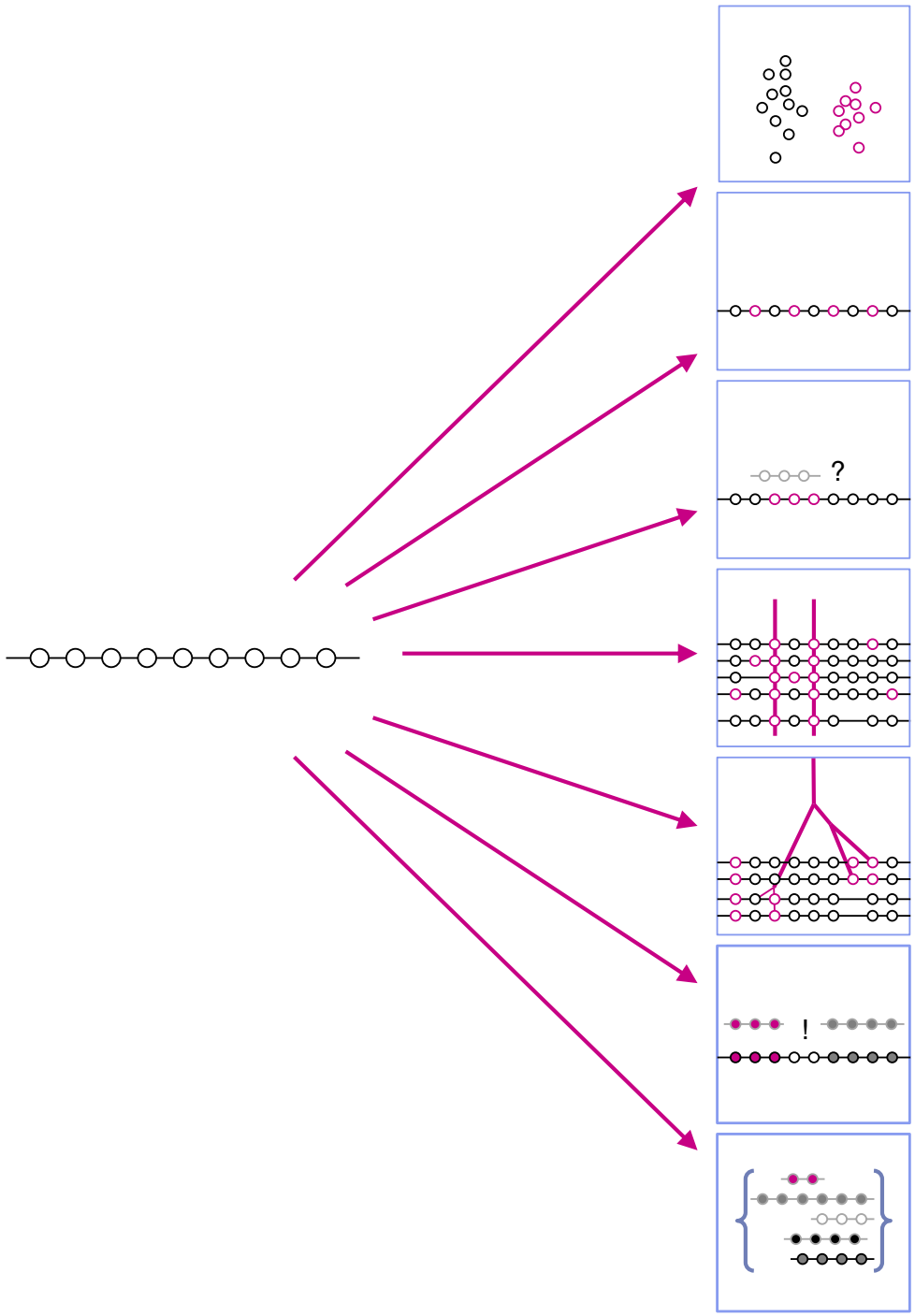
Juliet: "What's in a name? That which we call a rose/By any other name would smell as sweet." *Romeo and Juliet (II, ii, 1-2)*

Analysis by composition:	Analysis by comparison:	Analysis of conservation:
Count components, sum properties.	Compare patterns with an annotated dictionary.	Identify relatedness, analyze change.

Interestingly, the majority of information about sequences is not IN the sequence itself. It follows from the interaction of a biomolecule with its *context* and is discovered through *context-aware* analysis at various levels.

Dimensions of context:

- Composition
- Concatenation
- Comparison
- Correspondence
- Conservation
- Context
- Collaboration



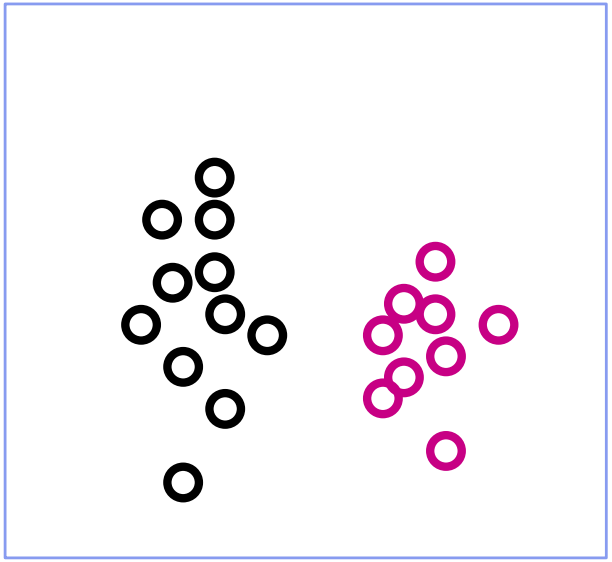
Composition

Unordered sets of amino acids.

Hypothesis: The weighted sum of properties of individual amino acids determines the aggregate properties of the protein.

Examples:

- molecular weight, isoelectric point, coefficient of extinction, abnormal composition



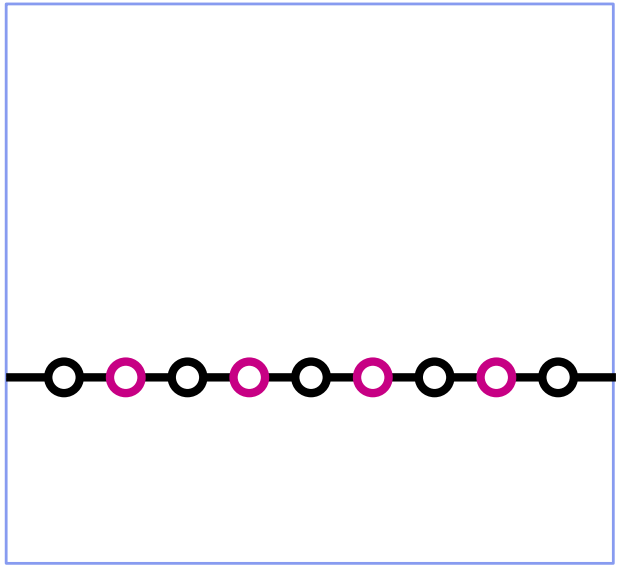
Concatenation

Amino acids in sequence context.

Hypothesis: The sequential arrangement of amino acids determines the properties of the protein.

Examples:

- Amphipathic 2° -structure
- Disordered segments
- Polyproline structures
- Transmembrane segments
- Salt-bridges
- Structure prediction



Comparison

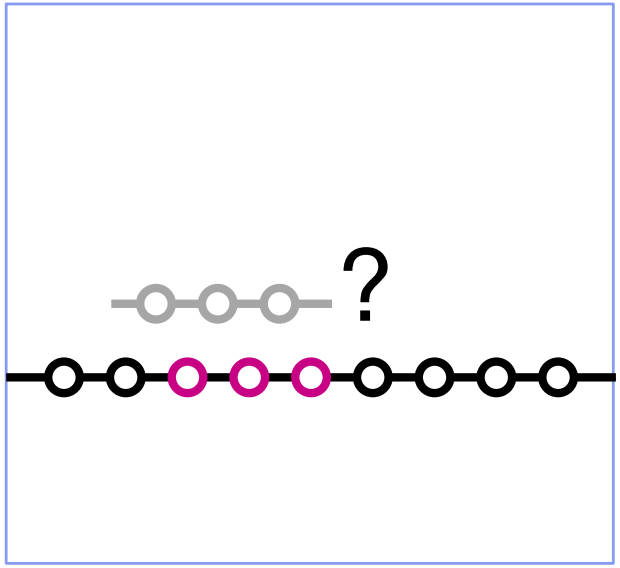
Comparison of a sequence with (abstract) patterns.

Hypothesis: similarity of sequence patterns to known patterns generates a similar function.

Pattern matching

Examples:

- Restriction sites
- Promoters & Operators
- Coiled coil domains
- Signal peptide cleavage sites
- Secondary structure
- Structure threading



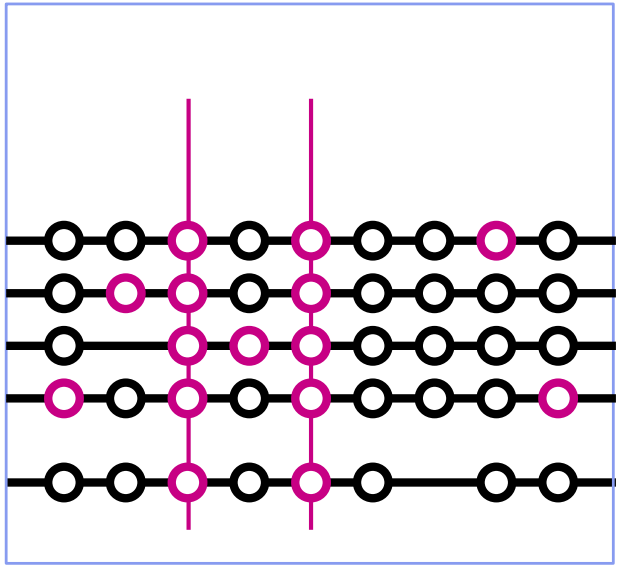
Correspondence

Analysis of corresponding positions in a sequence with other sequences.

Hypotheses: Average properties and property distributions are more informative than individual properties.

Examples:

- Pairwise and multiple alignment
- BLAST
- Structural superposition
- Structural motifs



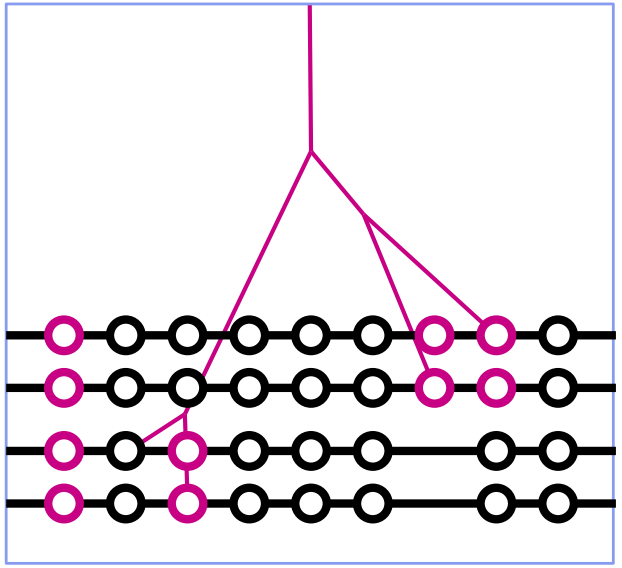
Conservation

Change of a sequence over time.

Hypotheses: Important features are conserved
Homologous sequences always have similar structure.
Homologous sequences usually have similar function.

Examples:

- Annotation transfer
- Functional change
- Phylogeny
- Analysis of evolutionary pressure



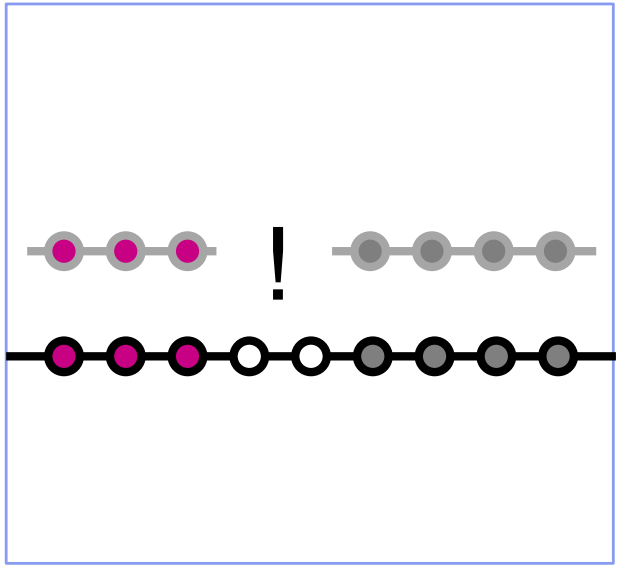
Context

Arrangement of functional elements.

Hypothesis: Proteins and protein domains that form complexes have functions in which each component acts in the context of the other.

Domain annotation

CDART

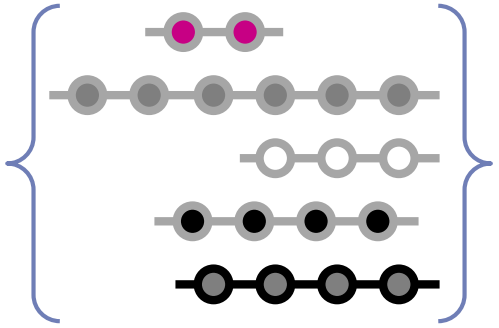


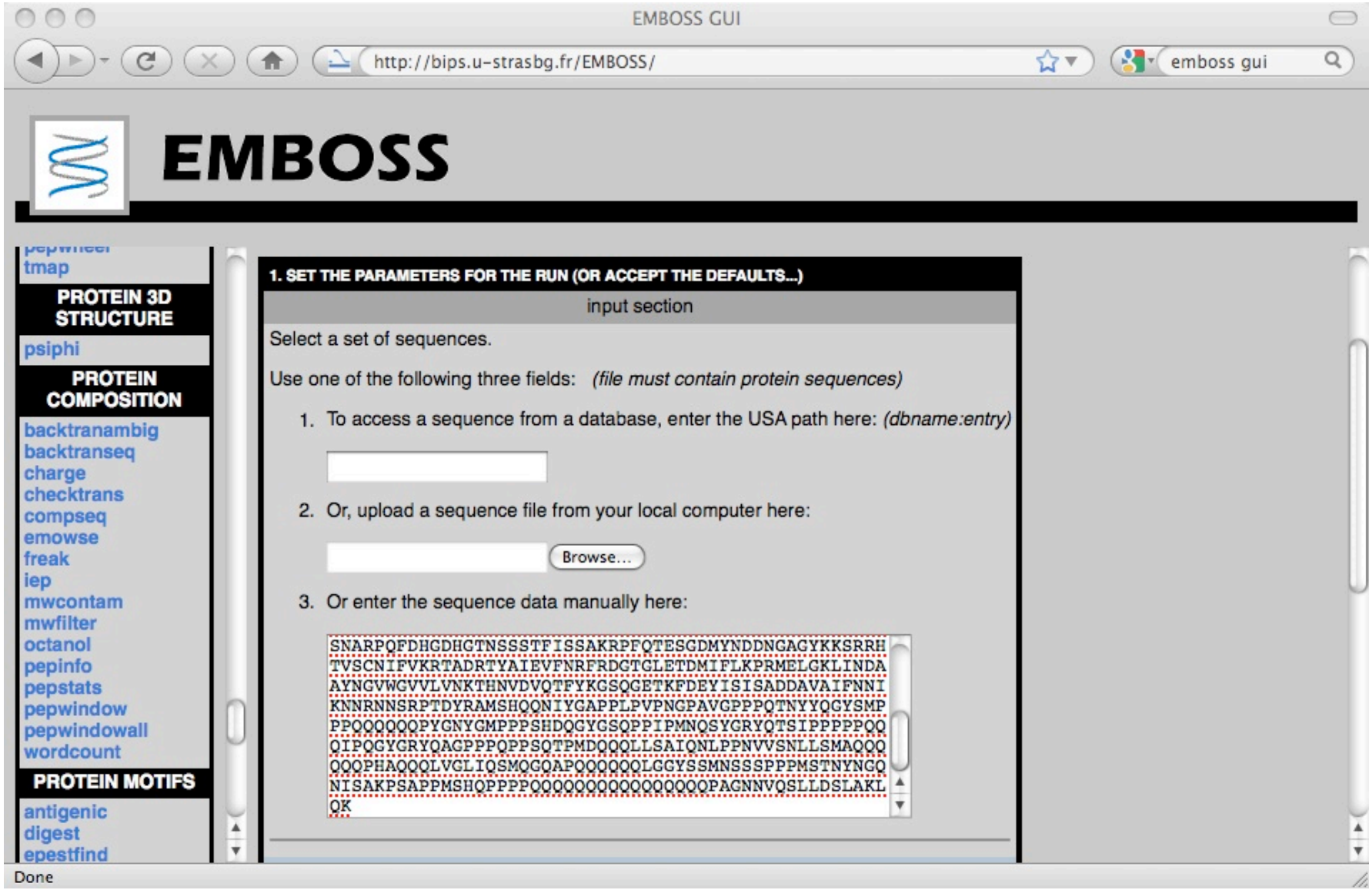
Collaboration

Collection of sequences from shared context.

Hypothesis: sequences that share a context also share a functional relationship.

- Co-expression / Co-regulation
- Co-location
- GO-term enrichment
- Metabolic / regulatory pathway
- Operon
- Phylogenetic footprints
- Systems biology




EMBOSS GUI
http://bips.u-strasbg.fr/EMBOSS/
emboss gui
EMBOSS
tmap
PROTEIN 3D STRUCTURE
psiphi
PROTEIN COMPOSITION
backtranambig
backtranseq
charge
checktrans
compseq
emowse
freak
iep
mwcontam
mwfilter
octanol
pepinfo
pepstats
pepwindow
pepwindowall
wordcount
PROTEIN MOTIFS
antigenic
digest
epestfind
1. SET THE PARAMETERS FOR THE RUN (OR ACCEPT THE DEFAULTS...)
input section
Select a set of sequences.
Use one of the following three fields: (file must contain protein sequences)
1. To access a sequence from a database, enter the USA path here: (dbname:entry)
2. Or, upload a sequence file from your local computer here:
Browse...
3. Or enter the sequence data manually here:
SNARPQFDHGDHGTNSSSTFISSAKRPFQTESGDMYNDDNGAGYKKSRRH
TVSCNIFVKRTADRTYAIEVFNRFRDGTGLETDMIFLKPRMELGKLINDA
AYNGVWGVVLVNKTHNVDVQTFYKGSQGETKFDEYISISADDAVAIFNNI
KNNRNNSRPTDYRAMSHQQNIYGAPPLPVPNGPAVGPPPQTNYYQGYSMP
PPQQQQQQPYGNYGMPPPSHDQGYGSQPPIPMNQSYGRYQTSIPPPPPQQ
QIPQGYGRYQAGPPPQPPSQTPMDQQQLLSAIQNLPPNVVSNLLSMAQQQ
QQQPHAQQQLVGLIQSMQGQAPQQQQQQLGGYSSMNSSSPPPMSTNYNGQ
NISAKPSAPPMSHQPPPPQQQQQQQQQQQQQQQQPAGNNVQSLLDSLAKL
QK
Done

```
PEPSTATS of Swi4.fa from 1 to 1093

Molecular weight = 123805.94            Residues = 1093
Average Residue Weight  = 113.272       Charge   = 23.5
Isoelectric Point = 9.2008
A280 Molar Extinction Coefficient  = 58600
A280 Extinction Coefficient 1mg/ml = 0.47
Improbability of expression in inclusion bodies = 0.952

Residue         Number          Mole%
A = Ala         41              3.751
B = Asx         0               0.000
C = Cys         7               0.640
D = Asp         54              4.941
E = Glu         56              5.124
F = Phe         33              3.019
[...]
```

(PEPSTATS: see EMBOSS package; EMBOSS GUI)

In R:

seqinr package (Excerpts)

cai	Codon Adaptation Index
computePI	Theoretical Isoelectric Point
count	Composition of dimer/trimer/etc oligomers
dotPlot	Dot Plot Comparison of two sequences
GC	fractional G+C content
pmw	Protein Molecular Weight
ucoweight	Weight of each synonymous codon
zscore	over- and under- representation of dinucleotides

Biostrings package (Excerpts)

alphabetFrequency	Calculate the frequency of letters in a biological sequence
countPattern	String searching functions
dinucleotideFrequency	Frequency of dinucleotides
findPalindromes	Searching a sequence for palindromes
longestConsecutive	Length of the longest substring containing only 'letter'
matchPDict	Matching a dictionary of patterns against a reference
pairwiseAlignment	Optimal Pairwise Alignment
reverseComplement	Reversing and complementing

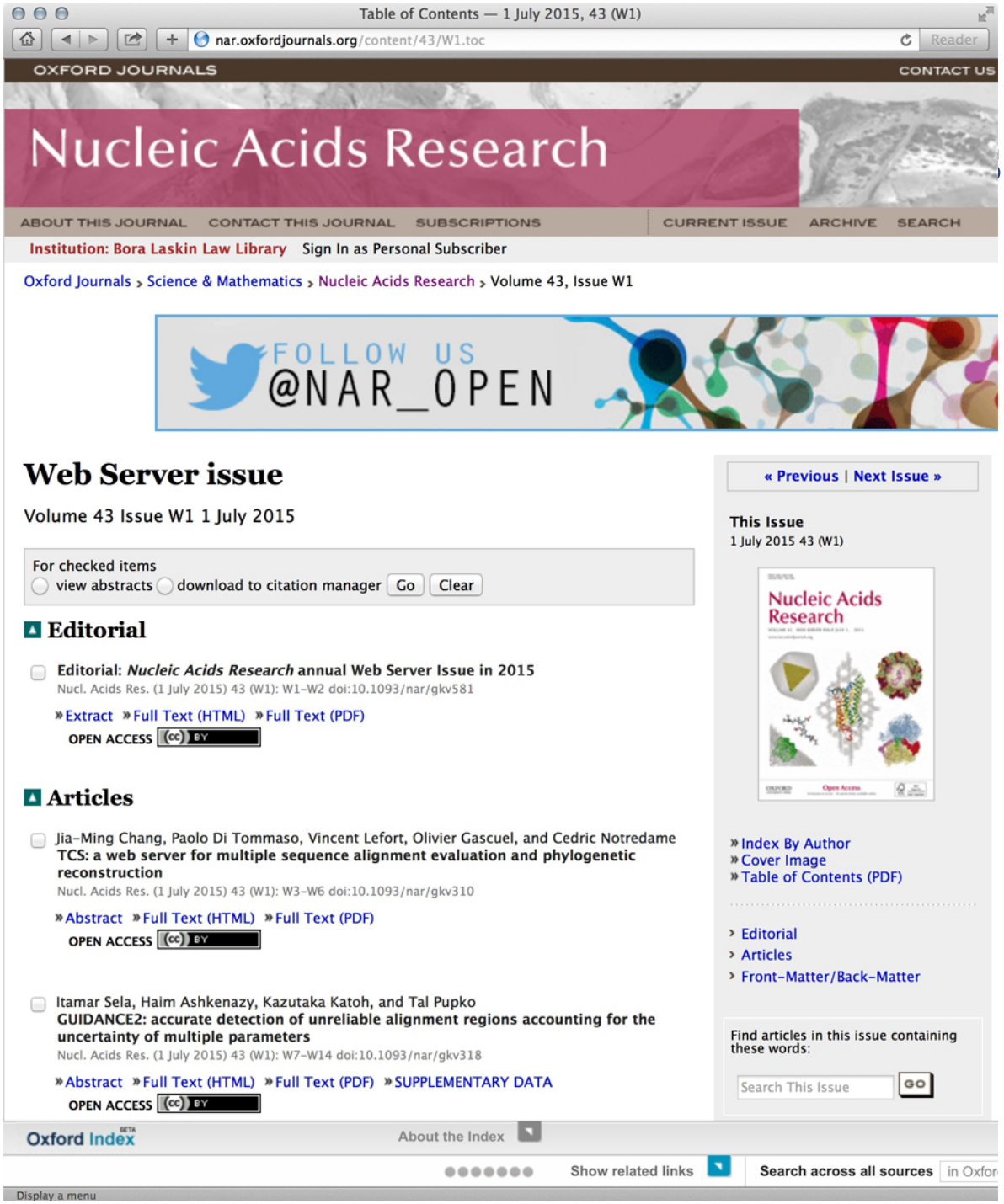


Table of Contents — 1 July 2015, 43 (W1)
nar.oxfordjournals.org/content/43/W1.toc
Reader
OXFORD JOURNALS
CONTACT US
Nucleic Acids Research
ABOUT THIS JOURNAL
CONTACT THIS JOURNAL
SUBSCRIPTIONS
CURRENT ISSUE
ARCHIVE
SEARCH
Institution: Bora Laskin Law Library
Sign In as Personal Subscriber
Oxford Journals › Science & Mathematics › Nucleic Acids Research › Volume 43, Issue W1
FOLLOW US
@NAR_OPEN
Web Server issue
Volume 43 Issue W1 1 July 2015
« Previous | Next Issue »
This Issue
1 July 2015 43 (W1)
For checked items
view abstracts
download to citation manager
Go
Clear
Editorial
Editorial: Nucleic Acids Research annual Web Server Issue in 2015
Nucl. Acids Res. (1 July 2015) 43 (W1): W1–W2 doi:10.1093/nar/gkv581
»Extract »Full Text (HTML) »Full Text (PDF)
OPEN ACCESS
Articles
Jia-Ming Chang, Paolo Di Tommaso, Vincent Lefort, Olivier Gascuel, and Cedric Notredame
TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction
Nucl. Acids Res. (1 July 2015) 43 (W1): W3–W6 doi:10.1093/nar/gkv310
»Abstract »Full Text (HTML) »Full Text (PDF)
OPEN ACCESS
Itamar Sela, Haim Ashkenazy, Kazutaka Katoh, and Tal Pupko
GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters
Nucl. Acids Res. (1 July 2015) 43 (W1): W7–W14 doi:10.1093/nar/gkv318
»Abstract »Full Text (HTML) »Full Text (PDF) »SUPPLEMENTARY DATA
OPEN ACCESS
»Index By Author
»Cover Image
»Table of Contents (PDF)
› Editorial
› Articles
› Front-Matter/Back-Matter
Find articles in this issue containing these words:
Search This Issue
GO
Oxford Index
About the Index
Show related links
Search across all sources
Display a menu

Given so many resources, the question is not:

What can you do?

But:

What should you do?

Never execute a procedure just because you can.
First clarify your objectives.
Then ask if the procedure is right for you:

- When is it appropriate?
- What data does it require?
- How is it used correctly?
- How are the results interpreted?
- How do the results support your objectives?

When to analyze: getting clear on the *workflow.*

Sample objectives:

Predict expressed sequence from genome sequence;

Identify functional residues in order to predict effects of sequence variation and correlate this with observed phenotypes;

Predict molecular weight, extinction coefficient to interpret experimental results;

Predict post-translational modification to provide hypotheses for evaluating experiments;

Predict domain boundaries to clone and express domains separately;

Annotate homologues to define evolutionary relationships;

Identify sets of co-expressed genes to predict genes of related function;

Predict interaction partners in order to deconstruct developmental / regulatory / metabolic pathways and identify drug targets.

When to analyze: getting clear on the *workflow.*

Sample objectives:

- Predict expressed sequence from ge
- Identify functional residues in order
  correlate this with observed phenoty
- Predict molecular weight, extinction
- Predict post-translational modificati
  experiments;
- Predict domain boundaries to clone
- Annotate homologues to define evolutionary relationships;
- Identify sets of co-expressed genes to predict genes of related function;
- Predict interaction partners in order to deconstruct developmental / regulatory / metabolic pathways and identify drug targets.

Generally, experiments should be able to lead to *predictions*. Usually these predictions are part of a larger workflow.

When to analyze: getting clear on the *workflow.*

Sample objectives:

Predict expressed sequence from genome sequence;

Identify functional residues in order to predict effects of sequence variation and correlate this with observed phenotypes;

Predict molecular weight, extinction coefficient to interpret experimental results;

Predict post-translational modification to provide hypotheses for evaluating experiments;

Predict domain boundaries to clone and exp

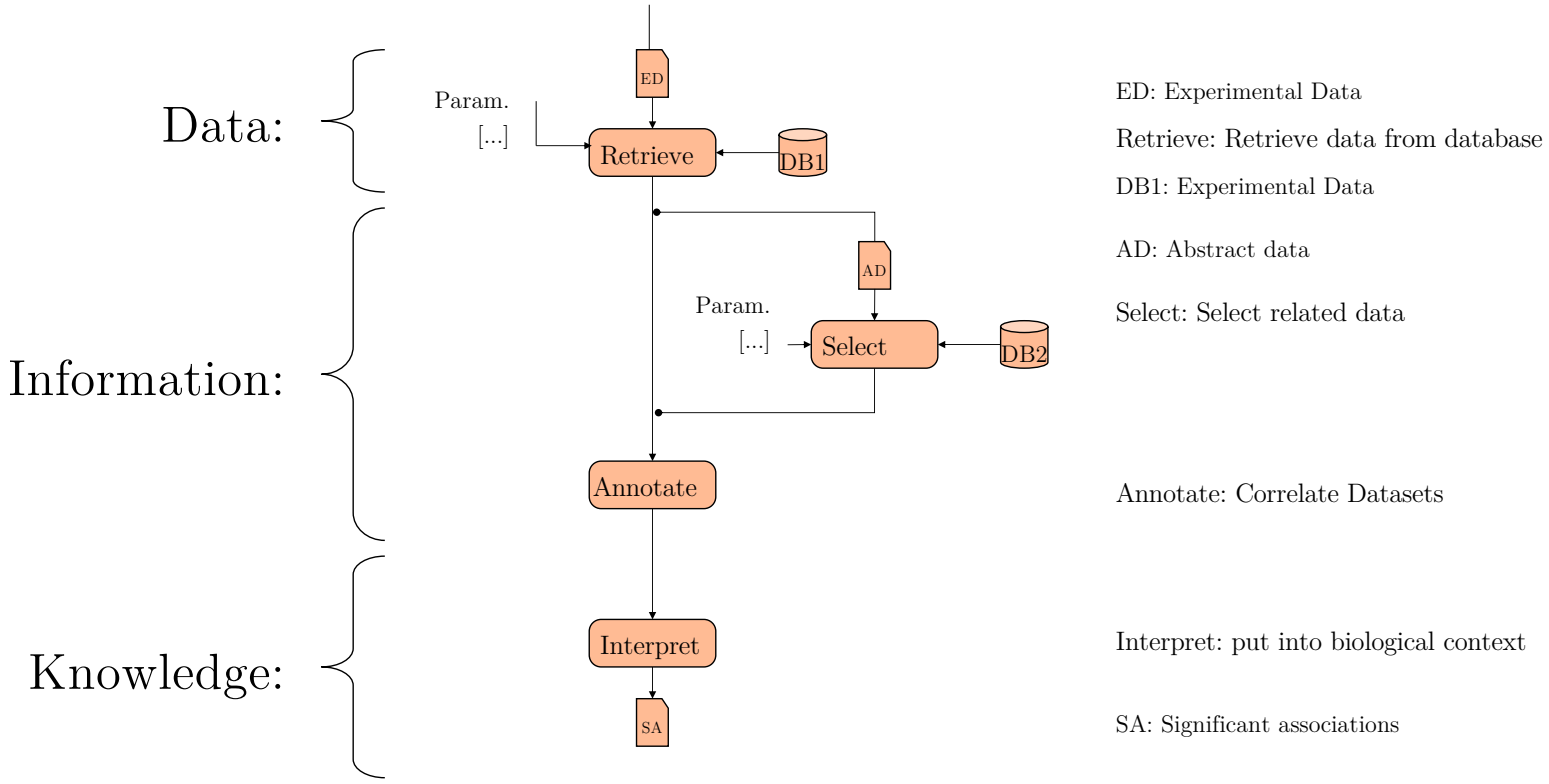
Annotate homologues to define evolutionary

Identify sets of co-expresse

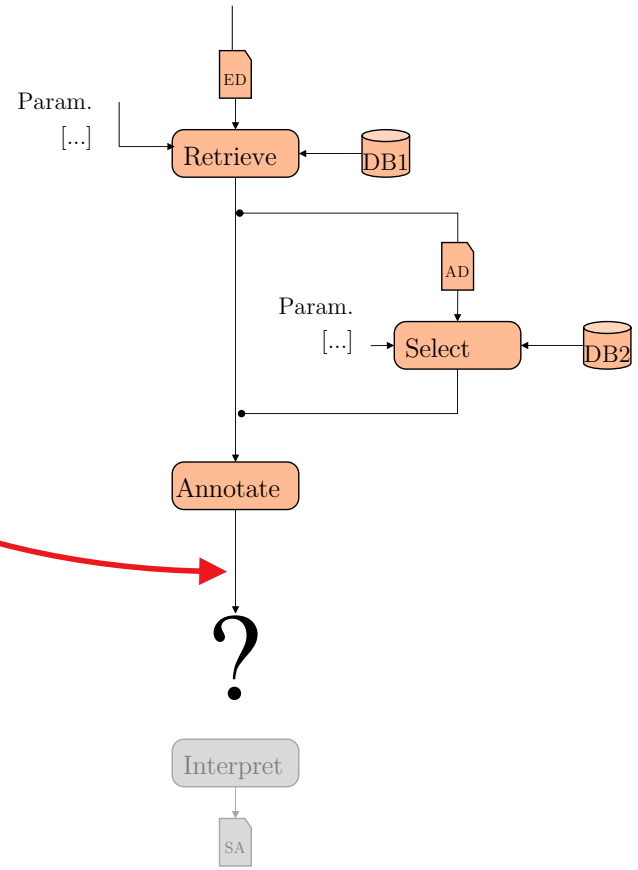
Predict interaction partners in orde to dec
metabolic pathways and identify drug targe

Sometimes (rarely), no experimental validation is possible. Then the prediction is the endpoint of the workflow and may lead to a new hypothesis ("Discovery Science").

Mapping data / information / knowledge into tasks. In general ...



Analysis too often stops here !



To complete the workflow, we need well defined, *integrated* processes that span the entire workflow from the primary data, to the ultimate *goals* of the experiment.

Examples:
Explanation / Prediction / Intervention

- Validated genetic markers
- Mechanistic insights into biological systems
- Understanding our phylogeny
- Rational protein engineering
- Rapid vaccine development
- ...

Goals are needed to *guide* processes !

For ...		Use ...
Untranslated regions, splice sites, regulatory regions, gene context, recent evolutionary variation ...	→	Finished genome sequence: chromosomes, contigs ... (Not ESTs, STS ...)
Translated, spliced nucleotide sequence, coding SNPs, isoforms ...	→	mRNA
All protein related questions ...	→	Protein sequence
Whenever possible:	→	Use refseq or SwissProt

Always be clear about taxonomy: Organism and strain!

For ...	Use ...
Untranslated regions, splice sites, regulatory regions, gene context, recent evolutionary	→ Finished genome sequence: chromosomes, contigs ... (Not ESTs, STS, ...)
Whenever possible:	→ Use refseq or Swissprot

Principle:

Use the sequence in which the information you are looking for is *conserved* in biology.

Always be clear about taxonomy: Organism and strain!

Conserved ...: Contribution to fitness function demonstrates biological relevance.

Treat sequence analysis like a wet-lab experiment:

Record essential parameters:

For *static parameters*, a link or reference may be sufficient.

Dynamic parameters need to be recorded completely.

(If possible).

Include controls!

Good choice for *negative controls*: shuffled sequences.

Positive controls: literature.

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

B O R I S . S T E I P E @ U T O R O N T O . C A

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