Bioinformatics of sequence and structure

- Sequence alignment
- Database searching
- Significance, e-values
- Phylogeny
- Gene ontology
- Homology modeling
- Drug Design
BLOSUM62: protein substitution matrix

Protein substitution scores are derived from statistics of substitutions in “Gold Standard” alignments.
Protein alignments are optimal paths through a matrix filled with substitution scores.
Protein alignments are optimal paths through a matrix filled with substitution scores.

ACTLIGYDKEEKWQMVV-SRRDI-ACTIK-LLGFENS-WQMILSKNNIC
Well aligned sequences express the history of mutations.

The tree topology is based on the full sequences.

Parsimony analysis determines the ancestral sequence characters.
substitution counts are summed from confidently aligned sequences
Counts become log-likelihood ratios

$$\text{LLR} = 2 \log_2(\text{observed}/\text{expected})$$

MSAs

BLOSUM62: protein substitution matrix

Protein substitution scores are derived from statistics of substitutions in “Gold Standard” alignments.

LLRs
Protein versus DNA alignments

Are protein alignment better?

• Protein alphabet = 20, DNA alphabet = 4.
  – Protein alignment is more informative
  – Less chance of homoplasy with proteins.
  – Homology detectable at greater edit distance
  – Protein alignment more informative

• Better Gold Standard alignments are available for proteins.
  – Better statistics from G.S. alignments.

• On the other hand, DNA alignments are more sensitive to short evolutionary distances.
Bioinformatics

- Sequence alignment
- **Database searching**
- Significance, e-values
- Phylogeny
- Gene ontology
- Homology modeling
- Drug Design
Database searching

Why do a database search?


**Pathology, epidemiology, ecology**: Determination of species, strain, lineage, phylogeny.

**Biophysics**: Prediction of RNA or protein structure, effect of mutation.

one sequence \(\rightarrow\) GenBank, PIR, Swissprot, GenEMBL, DDBJ \(\rightarrow\) lots of sequences
Fast Database Searching

**BLAST**  
S. Altschul *et al.*

First make a set of lookup tables for all 3-letter (protein) or 11-letter (DNA) matches.

Make another lookup table: the **locations of all 3-letter words** in the database.

Start with a match, extend to the left and right until the score no longer increases.

Very fast. Selective, but not as sensitive as slower search methods (SSEARCH). Reliable statistics. Heuristic, not optimal.
Each 3-tuple is scored against all 8000 possible 3-tuples using BLOSUM. The top scoring 50 are kept as that 3tuple’s “neighborhood words”
All locations of query words in targets sequences are pre-calculated! So it is fast.
• FASTA algorithm strings together HSPs to form a sequential pathway.
• Only the highest scoring targets pass to
Results of BLAST can be expressed as a sequence alignment (COBALT), or a tree. BLAST annotations can be used to extract functional annotations from GO. Coordinates for known structures can be extracted from PDB.
## forms of BLAST

<table>
<thead>
<tr>
<th>BLAST</th>
<th>query</th>
<th>database</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
</tr>
<tr>
<td>blastp</td>
<td>protein</td>
<td>protein</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein</td>
<td>translated DNA</td>
</tr>
<tr>
<td>blastx</td>
<td>translated DNA</td>
<td>protein</td>
</tr>
<tr>
<td>tblastx</td>
<td>translated DNA</td>
<td>translated DNA</td>
</tr>
<tr>
<td>psi-blast</td>
<td>profile</td>
<td>protein</td>
</tr>
<tr>
<td>phi-blast</td>
<td>protein+pattern</td>
<td>protein</td>
</tr>
<tr>
<td>rps-blast</td>
<td>protein</td>
<td>profile</td>
</tr>
</tbody>
</table>
Bioinformatics

• Sequence alignment
• Database searching
• **Significance, e-values**
• Phylogeny
• Gene ontology
• Homology modeling
• Drug Design
How significant is that?

...a specific inference. Thanks.

...as opposed to...

...how likely the data would not have been the result of chance,...

Please give me a number for...
Dayhoff's randomization experiment

Aligned scrambled Protein A versus scrambled Protein B 100 times (re-scrambling each time).

NOTE: scrambling does not change the AA composition!

Results: A Normal Distribution

The significance of a score is measured as the probability of getting this score in a random alignment.
Lippman's randomization experiment

Aligned Protein A to 100 natural sequences, not scrambled.

Results: A wider normal distribution (Std dev = ~3 times larger)

WHY? Because natural sequences are different than random.

Even unrelated sequences have similar local patterns, and uneven amino acid composition.

Lippman got a similar result if he randomized the sequences by words instead of letters.

Was the significance over-estimated using Dayhoff's method?
Why was Lippman’s distribution wider?

- *low complexity* sequences score higher (and lower) by chance than scrambled.

```
AAA.AAAAAHAAAAAA.AAAK.AAAAAAEAA
AAADAAAAA.AAAAALAA.GGAAAAA.AA
```

- Local sequence motifs. Protein sequences are made up of recurrent *words*.

```
U...NRELATEDMEANING
.MEANRELATIVE... .
```

*False scores are not random scores.*
$P(S > x)$

E(M) gives us the expected length of the longest number of matches in a row. But, what we really want is the answer to this question:

How good is the score $x$? (i.e. how significant)

So, we need to model the whole distribution of chance scores, then ask how likely is it that my score or greater comes from that model.
Suppose you had a Gaussian distribution “dart-board”. You throw 1000 darts randomly. Score your darts according the number on the X-axis where it lands. What is the probability distribution of scores?
Answer: The same Gaussian distribution! (duh)
What if we throw 10 darts at a time and keep only the highest-scoring dart (extreme value)? What is the distribution of the extreme values?
False alignment scores fit the extreme value distribution

\[ P(x) = \exp(-x - e^{-\lambda(x-u)}) \]
Altschul's randomization experiment

- Generate optimal false alignment scores using DP.
- Integrate and take the log of the EVD* to make it a linear equation: \( \log(P(S \geq x)) \approx \log(K_{mn}) - \lambda x \)
- Plot \( \log(S \geq x) \) versus \( x \) to get the parameters of the EVD.
- Significance is correctly estimated by the EVD.

\[
\text{slope} = -\lambda, \\
\text{intercept} = \log(K_{mn})
\]

* \( P(x) = \exp(-x - e^{\log(K_{mn})-\lambda x}) \)
e-values in BLAST

• Every BLAST "hit" produces a bit-score.

• Parameters (pre-set) for the EVD are used to get the e-value.

• The e-value is defined as the expected number of times this score will occur by chance.

\[ e\text{-value} = P(S \geq x) \times m \]

where x is the bit-score, m is the size of the database, and \( P(S \geq x) \) is the value of the integrated EVD at x.
History of significance in database searching

• Significance of a score is measured by the probability of getting that score by chance.

• History of modeling “chance” in alignments
  • 1970’s Dayhoff: Guassian fit to scrambled alignments
  • 1980’s Lipman: Gaussian fit to false alignments
  • 1990’s Altschul: EVD fit to optimal false alignments
Pop-quiz

You did a BLAST search using a sequence that has absolutely no homologs in the database. Absolutely none.

The BLAST search gave you false “hits” with the top e-values ranging from 0 to 20. You look at them and you notice a pattern in the e-values.

How many of your hits have e-value $\leq 10$?
A story about faking significance

Correlation

\[ r = \frac{\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sum_i (x_i - \langle x \rangle)^2 \sum_i (y_i - \langle y \rangle)^2} \]

Pearson's correlations coefficient.

Or

Pearson's product moment correlation
Non-linearity is not picked up by correlation

All of these examples have $r=0.816$
Re-sampling will fix some of these.
Correlation Confidence by Resampling

• Start with paired data (x,y), calculate $r$. Let’s say $r=0.511$

• Randomly associate x and y values.

• Calculate $r_{ran}$

• Repeat 10000 times.

• Significance is $p =$ number of times $r_{ran} > 0.511$, divided by 10000. (If $r$ is negative, count $r_{ran} < r$)
Resampling example

The y-values have been randomly swapped, 10000 times. 10% of the time, r is $\geq 0.816$, therefore $p$-value=$0.10$.
Correlation using Boolean data

True is assigned 1, False 0.

\[
\begin{align*}
\sum_{i} (x_i - <x>) (y_i - <y>) \\
\sum_{i} (x_i - <x>)^2 \sum_{i} (y_i - <y>)^2
\end{align*}
\]

\[
\rho = \frac{\sum_{i} (x_i - <x>) (y_i - <y>)}{\sqrt{\sum_{i} (x_i - <x>)^2 \sum_{i} (y_i - <y>)^2}}
\]

\[
<x> = \frac{4}{12} = 0.33
\]

\[
<y> = \frac{6}{12} = 0.50
\]

\[
\rho = 0.71
\]

\[
p = 0.025
\]
**Discussion Topic : The trial of Merck Smith-Kline**

Merck Smith-Kline was the author on a study of Trioxx, an anti-inflammatory drug used to treat arthritis, for which it was know to be effective. The study followed over 500 long-time Trioxx users and an equal number of control subjects who had never used the drug. Dr. Smith-Kline was looking for correlations between the use of Trioxx and the incidence of any disease other than arthritis, in any demographic group. He noted in the study that Tunisian Americans, in the age range from 45-55, male or female, and who had been a vegetarian for more than 6 months at any time in their lives, had a "strong negative correlation" between the use of Trioxx and the incidence of restless leg syndrome (RLS), and began touting Trioxx as an effective anti-RLS drug.

The numbers were as follows:

<table>
<thead>
<tr>
<th></th>
<th>Trioxx non-users</th>
<th>Trioxx users</th>
</tr>
</thead>
<tbody>
<tr>
<td>RLS</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>no RLS</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

Dr. Smith-Kline correctly calculated the correlation* between Trioxx and RLS as follows:

\[
\text{Corr} = \frac{\sum (u_i - \langle u \rangle)(r_i - \langle r \rangle)}{\sqrt{\sum (u_i - \langle u \rangle)^2 \sum (r_i - \langle r \rangle)^2}}
\]

The sums were carried out over all 32 subjects in the subset, and the resulting correlation was -0.378. This confidence level was cited as 99%, since the p-value for this correlation was 0.01, The sample size of 32 and the uneven distribution of subjects with RLS were taken into account.

The data itself was collected correctly. The calculations were correct, both for the correlation and for the confidence. Yet Merck Smith-Kline did something dishonest in this study. What was it and what specific question would you ask him to reveal his dishonesty?
Bioinformatics

- Sequence alignment
- Database searching
- Significance, e-values
- Phylogeny
- Gene ontology
- Homology modeling
- Drug Design
Evolutionary time

Cladogram

Phylogram

Ultrametric tree

parenthesis (notation can have both labels and distances).
A multiple sequence alignment is made using many pairwise sequence alignments.
Multiple sequence alignment

1. align all pairs
2. pairwise align two most similar first
3. align next most similar
4. repeat until all sequences are aligned

\[ S(P,[W,F]) = \frac{1}{2}(S(P,W) + S(P,F)) \]
Construct a distance-based tree

Draw tree here
• Start with unrooted tree, using Neighbor joining.
• choose root to get guide tree
• progressive alignment
  – matches are scored using sequence weights
  – gaps are position dependent, as follows
    • GOP* lower for polar residues
    • GOP zero where there is already a gap

*gap opening penalty
Figure 3. The variation in local gap opening penalty is plotted for a section of alignment. The initial gap opening penalty is indicated by a dotted line. Two hydrophilic stretches are underlined. The lowest penalties correspond to the ends of the alignment, the hydrophilic stretches and the two positions with gaps. The highest values are within 8 residues of the two gap positions. The rest of the variation is caused by the residue specific gap penalties (12).
Parsimony: Finding the tree with the minimum number of mutations

Given a tree and a set of taxa, one-letter each (1) choose optional characters for each ancestor. (2) Select the root character that minimizes the number of mutations by selecting each and propagating it through the tree.

minimum 2 mutations

minimum 1 mutation
Columns in a MSA have a common evolutionary history

By aligning the sequences, we assert that the aligned residues in each column had a common ancestor.
Orthologs/paralogs

Orthologs: homologs originating from a speciation event
Paralogs: homologs originating from a gene duplication event.
How do I know it’s a paralog?

• If it’s a paralog, then at some point in evolutionary history, a species existed with two identical genes in it.
  • One may have been lost since then. (Descendants are still paralogs!)
  • Paralogs can be from different species.
• Paralogous genes have more than the expected sequence divergence.
  • Because they are more likely to have different functions
  • Because they diverged earlier than the speciation event.
• Without species information or functional information, it’s impossible to tell
“Boot strap analysis”

• A method to validate a phylogenetic tree, branchpoint by branchpoint.

• Requires a means to generate independent trees. (For example trees generated from different regions of the mitochondrial genome.)

• Choose the representative tree as the ‘parent’. Calculate the following:

  For each branchpoint in the parent tree,
  For each tree, ask
    Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)
  Bootstrap value = number of trees having the branchpoint / total trees.
Comparing branchpoints

For each branchpoint in the parent tree,
For each tree, ask
Is there a branchpoint having the same subclade contents (i.e. same taxa, any order)
Bootstrap value = number of trees having the branchpoint / total trees.

\[ P((A,B),C) = \frac{5}{8} \]
Bioinformatics

• Sequence alignment
• Database searching
• Significance, e-values
• Phylogeny
• **Gene ontology**
• Homology modeling
• Drug Design
Ontology

• Ontologies relate **facts** to **knowledge**
• facts
  – may be known/unknown/little known
  – not attached to knowers
  – unchanging
• Knowledge
  – attached to knower
  – may disappear
Gene Ontology

- Gene annotation system
- Controlled vocabulary that can be applied to all organisms
- Used to describe gene products
What is the Gene Ontology?

A (part of the) solution:

- A controlled vocabulary that can be applied to all organisms

- Used to describe gene products - proteins and RNA - in any organism
GO: Three ontologies

What does it do?

What processes is it involved in?

Where does it act?

Molecular Function

Biological Process

Cellular Component

gene product
Cellular Component

- where a gene product acts
Cellular Component

Mitochondrial membrane

cellular_component
0(+7268) genes

cell
1(+5434) genes

intracellular
46(+4974) genes

organelle
0(+3919) genes

membrane-bound
organelle
0(+3600) genes

membrane
83(+942) genes

cyttoplasm
311(+2556) genes

intracellular
organelle
0(+3919) genes

membrane-bound
organelle
0(+3600) genes

organelle
0(+572) genes

mitochondrion
715(+382) genes

mitochondrial
membrane
16(+194) genes
Biological Process
Gluconeogenesis
Molecular Function

- A single reaction or activity, not a gene product
- A gene product may have several functions
- Sets of functions make up a biological process
Molecular Function

hexose kinase
What can you do with GO?

• Access gene product functional information

• Find how much of a proteome is involved in a given process, function, or component in the cell

• Map GO terms to annotate databases

• Provide a link between biological knowledge and …
  • gene expression profiles
  • proteomics data
Selecting microarray subsets based on GO reveals drug target

Figure modified from http://en.wikipedia.org/wiki/Image:Microarray-schema.jpg

courtesy of Shabana Shabeer, Albert Einsteing School of Medicine
OBO

• Open Biological and Biomedical Ontologies
• Ontologies are becoming organized, follow standard formats, are related to each other.
• http://www.obofoundry.org/
Bioinformatics

• Sequence alignment
• Database searching
• Significance, e-values
• Phylogeny
• Gene ontology
• **Homology modeling**
• Drug Design
Homology modeling
also called “comparative modeling”

• Sequences that have similar sequence have similar structure.
• Therefore we can model a sequence of unknown structure based on a homolog of known structure.
• Differences between homolog functions can be inferred.

Percent sequence identity

90% 50% 25% 5%
trivial useful difficult forget it.
Global positioning of SSEs is conserved.
Sidechain locations are conserved.
Sidechain orientation (rotamers) are conserved.
Why do homology modeling?

• Predict functional differences.
• Predict binding sites.
• Predict mutational effects.
• Predict drug interactions.
• Design all of the above.

• It’s easier than growing crystals.
What an alignment means to a biologist

Given this alignment...

Biologist infers...

- The gene was extended by one residue at the N-terminus.
- The Phe is conserved.
- Four residue deletion occurred between G to H.
- A non-similar mutation Y->K occurred.
- A two-residue insertion occurred between M and Q.
- A similar mutation F->Y occurred.

Aligned positions share a common ancestral position.
What an alignment means to a modeler

Given this alignment...

target

ACDEFG....HIKLMNPQRSTVWY

||:||    || :|  | ||||:

.CDDFGACDGHIYIM..QQSTVWF

template

• Modeler program should...
• Add Ala to the N-terminal Cys using energy minimization.
• Keep the conserved Phe sidechain and backbone.
• Cut out the four residue insertion and connect G to H.
• Switch non-similar sidechains Y->K. Possibly move backbone.
• Possibly pick another alignment.
• Cut at M-Q, insert two residues, Asn-Pro.
• Switch similar sidechains F->Y. Keep backbone fixed.

Aligned positions share a common spatial position.
In what order do I model?

Modeling philosophy: MAXIMIZE USE OF THE INFORMATION IN THE TEMPLATE

This means...

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIGH</td>
<td>Assign coordinates based on homology first. <strong>Identity</strong> first, then <strong>similarity</strong>. Then make <strong>deletions</strong>. Then add <strong>insertions</strong>.</td>
</tr>
<tr>
<td>LOW</td>
<td>Then build <strong>extensions</strong>.</td>
</tr>
</tbody>
</table>
Specifically, How?

What actions do I take to build a model?

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity</td>
<td>keep coordinates. no change.</td>
</tr>
<tr>
<td>similarity</td>
<td>mutate sidechain</td>
</tr>
<tr>
<td>deletions</td>
<td>remove residues, make new peptide bond, energy minimize.</td>
</tr>
<tr>
<td>insertions</td>
<td>predict loop conformation, position a loop. Make two new peptide bonds, energy minimize.</td>
</tr>
<tr>
<td>extensions</td>
<td>predict extension conformation, position an extension. Make new peptide bond, energy minimize.</td>
</tr>
</tbody>
</table>
Automated Loop Search

Loops of the right length in the database are superimposed on the anchor residues and the RMSD is calculated.

MOE keeps the loops with the best RMSDs to anchors, and lowest energy.

pre-flex anchor residues

post-flex anchor residues

gap distance

indel
No obvious way to make it better by energy minimization

Phi-Psi outlier found

Unalign 2 or 3 residues and run Homology Model — 2nd pass.
After 2nd pass Homology Model…

New loop has no phi-psi outliers.
Bioinformatics

• Sequence alignment
• Database searching
• Significance, e-values
• Phylogeny
• Gene ontology
• Homology modeling
• Drug Design
Prostaglandins

First found in prostate gland (seminal) fluid

20-carbon fatty acid + oxygen -->

cyclooxygenase

arachidonic acid

Prostaglandins
GPCRs

- G-protein coupled receptors
- Transmembrane, 7 helices
- ~800 genes in human
- ligand-specific signal transduction -- olfactory, light, neurotransmitters, inflammation
Effects of aspirin and other pain killers?

Aspirin works on both COX-1 and COX-2 to inhibit arachidonic acid’s entry into the active site of the enzyme --acetyl group of aspirin binds to serine in COX -- by blocking the activity of the COX enzymes, this relieves some of the effects of pain and fever --”nonselective” --many side effects

Tylenol—thought to have effects through inhibiting the activity of COX-3, an alternatively spliced form of COX-1
The difference between cox-1 and cox-2

<table>
<thead>
<tr>
<th>name</th>
<th>expression</th>
<th>function</th>
<th>Position 509</th>
</tr>
</thead>
<tbody>
<tr>
<td>cox-1</td>
<td>constitutive</td>
<td>housekeeping enzyme</td>
<td>Ile</td>
</tr>
<tr>
<td>cox-2</td>
<td>induced</td>
<td>associated with inflammation</td>
<td>Val</td>
</tr>
</tbody>
</table>

paralogous
65% identical

A drug that binds cox-2 but not cox-1 would specifically address inflammation without gastrointestinal side-effects.
Short History of COX-2 Inhibitors (2)

• mid 1990’s - X-ray crystal structures showed that COX-2 active site has a side pocket that is absent in COX-1.

• late 1990’s - inhibitors designed that preferentially bind COX-2, are equally effective as NSAIDs, but are reported to cause less GI damage.

• 1998/1999 - Celebrex and Vioxx approved by FDA for osteoarthritis and rheumatoid arthritis.

• 2000’s - next generation of drugs developed with even higher selectivity for COX-2 over COX-1.

Triumph for rational drug design
Rational Drug Design (1)
- Based on molecular and structural studies, informed by cell biology, physiology, pharmacology, etc.
Exercise 24

- Open PDB ID = 3ln1, COX-2 bound to celebrex (related to VIOXX)

- Delete all waters and all ligands except one copy of celebrex (CEL) and one COX-2 enzyme. Delete all extra copies of enzyme and CEL.

- Open PDB ID = 1ddx

- Delete all waters and all ligands except one copy of prostglandin (PGX) and one COX-2 enzyme. Delete all extra copies of enzyme and CEL.

- Superpose the two COX-2 structures, making sure to synchronize by Tag. Delete 1ddx protein chain.

- Show CEL and PGX with electrostatic surface of 3ln1.
A drug in action

Celebrex (celecoxib)
prostaglandin H2 (product)

Drug sterically blocks arachadonic acid binding, preventing enzyme activity.