Sequence analysis course

Lecture 8

Sequence databank searching 1
Sequence Databank Searching - Part 1

**Sequence**
- a string of characters that represents the chain of building blocks in a heteropolymer
- building blocks are amino acids (proteins) and nucleotides (DNA, RNA)

**Backbone - Side chain**
- backbone of proteins and poly-nucleotides is invariant!
- sidechains and bases make the difference: they define the sequence

**Database Searching**
Sequence databank searching is the process of extracting homologues of one or several query sequence(s) from a sequence database.
Backbone and sidechains
Protein main-chain and side-chain

http://www.chemistry.gatech.edu/faculty/williams/bCourse_Information/6521/protein/peptide_bond/down/six_coplaner_atoms.gif

The coloured area indicates planar configuration of enclosed atoms

No rotational freedom in within peptide bond plane, but rotational freedom at both C-alpha backbone bonds! NH-CO is usually 'trans' (as shown here) as opposed to 'cis'
# Nucleotides and Amino Acids

**Nucleotides**

<table>
<thead>
<tr>
<th>Letter</th>
<th>Nucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
</tbody>
</table>

**Amino acids**

<table>
<thead>
<tr>
<th>Letter</th>
<th>Amino Acid</th>
<th>Letter</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Ala Alanine</td>
<td>P</td>
<td>Pro Proline</td>
</tr>
<tr>
<td>C</td>
<td>Cys Cysteine</td>
<td>Q</td>
<td>Gln Glutamine</td>
</tr>
<tr>
<td>D</td>
<td>Asp Aspartatic acid</td>
<td>R</td>
<td>Arg Arg</td>
</tr>
<tr>
<td>E</td>
<td>Glu Glutamic acid</td>
<td>S</td>
<td>Ser Serine</td>
</tr>
<tr>
<td>F</td>
<td>Phe Phenylalanine</td>
<td>T</td>
<td>Thr Threonine</td>
</tr>
<tr>
<td>G</td>
<td>Gly Glycine</td>
<td>V</td>
<td>Val Valine</td>
</tr>
<tr>
<td>H</td>
<td>His Histidine</td>
<td>W</td>
<td>Trp Tryptophan</td>
</tr>
<tr>
<td>I</td>
<td>Ile Isoleucine</td>
<td>Y</td>
<td>Tyr Tyrosine</td>
</tr>
<tr>
<td>K</td>
<td>Lys Lysine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Leu Leucine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>Met Methionine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>Asn Asparagine</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sequence notation - Format Conventions

A sequence is composed of a name (often including an accession number) and the residue string. A sequence databank is a formatted (and often sorted) list of sequences (here FASTA format).

>lepi epidermal growth factor (Mus musculus)
NSYPGCPSSYDGYCLNGGVCMHIESLDSYTNCNVIGYSGDRCQTRDLRWWELR
>lixa EGF-like module coagulation factor (Homo sapiens)
VDGDQCESNPCLNGSCKDDINSYECWCPFGFEGKNCEL

Proteins are written from the N-terminus to the C-terminus:

\( (+) \ H_3N-CH(R1)-CO-NH-CH(R2)-CO-.(n).-NH-CH(Rn+2)-COO (-) \)

Nucleotide sequences are defined within a 'reading frame', because an amino acid is defined by a nucleotide triplet. The notation is from 5' to 3':

\( (-) \ OPO(O'-')(OR1)-OPO(O'-')(OR2)-.(n).-OPO(O'-')(ORn+2) \)
Sequence Notation - Name Conventions

The bond between two nucleotides is called a 'phosphodiester bond', and the molecule is called a 'dinucleotide'.

The bond between two amino acids is called 'peptide bond', which is chemically an amide bond, and the molecule is called a 'dipeptide'.

2 - dipeptide, e.g. GA or Gly-ALA or glycinyl-alanine
3 - tripeptide, e.g. YGA or Tyr-Gly-ALA or tyrosyl-glycinyl-alanine
4 - tetrapeptide
5 - pentapeptide
...
tens - oligopeptide
...
>~50 - polypeptide, protein

A protein adopts a folded structure with a hydrophobic core.
Sequence Notation - Positions and Chains

Mutation
Y35G-BPTI (bovine pancreatic trypsin inhibitor)
mutation from Tyr to Gly at position 35

K(B29)P-insulin
mutation from Lys to Pro at position 29 in chain B

Des(B27-B30)-insulin-B26-carboxamide
residues 27 to 30 deleted in chain B and C-terminus amidated

Chain notation
Chains are denoted A, B, C, D ... in successive order.

Disulfide bridges
Oxidation of proximous Cys Cys pairs leads to formation of a covalent
Cy-Cy disulfide bond (cystein bridge)
Hierarchical (self)organisation

Primary structure: sequence

Secondary structure: repetitive backbone angles -> repetitive 3D configuration
   alpha-(3.6,13)helix, beta-sheet (parallel, antiparallel)
   pi-(3.0,10)helix, beta-turn, gamma-turn, loop

Super-secondary structure: combinations of secondary structure elements
   helix-turn-helix, helix-turn-sheet

Domain: Packing of secondary structure elements around common core
   The domain is of central importance to protein evolution.
   It is a stable structural/functional entity with a core.
   Many proteins are composed of several domains.

Tertiary structure: total structure of one chain

Quaternary structure: association of several chains
Protein structure hierarchical levels

PRIMARY STRUCTURE (amino acid sequence)

VHLTPEEKSAVTALWGKVNVD
EVGGEALGRLLVVYPWTQRFF
ESFGDLSTPDAVMGNPKVKAH
GKKVLGAFA SDLAHLDNKGF
ATLSELHCDKLHVDPCENFRLLG
NVLCVLAHFKEFTPPVQAA
YQKV VAGVANAHKYH

SECONDARY STRUCTURE (helices, strands)

QUATERNARY STRUCTURE (oligomers)

TERTIARY STRUCTURE (fold)
Hemoglobin
Some Known Genomes

Eukariotes
Human, Rat, Mouse
Anopheles gambiae, Drosophila melanogaster, Arabidopsis thaliana, Oryza sativa, Solanum tuberosum

Parasites
Brugia malayi, Entamoeba histolytica, Plasmodium falciparum, Plasmodium yoelii, Toxoplasma gondii, Trypanosoma brucei, Trypanosoma cruzi, Schistosoma mansoni

Microbes
Chlamydia pneumoniae, Haemophilus influenzae, Helicobacter pylori, Mycobacterium tuberculosis, Mycoplasma genitalium, Vibrio cholerae

Fungal Projects
Cryptococcus neoformans, Aspergillus fumigatus

(http://www.tigr.org/tdb/)
(http://www.ensembl.org/)
Why are genome sequences so important?

The cellular machinery is extremely complex.

The transformation of genomic information from the cell to a text sequence is a reduction in complexity.

This formalisation makes genomic information accessible.

Compare to other formalisations:
- Carl Linnaeus (1707-1778): systematic classification
- Charles Darwin (1809-1982): Evolution
- Alan Turing (1912-1954): Turing machine
Reductionism

"Ceci n'est pas une pipe" (Rene Magritte).

This is not a protein.

>ENSANGP00000000001 Gene:ENSANGG00000000001 Status:novel
LDGSAVHPESYPVERILAKLEQTVDSLGLNSNLLLRTLKPADYTEDEQFGVPTVTDIIGEL
DKPGRDPRPEFKTATFKEGVEKISDLVPKVLELGVTNVTNFAGAFVPIGIVHQDGLVHSS
LTDVFKVDPREVVKAGDIVRVKVLVDVPDKRISLTMRLDEKAGQPARKPAEPRHTGNAK

Conclusion: Sequence-based tools use a reduced amount of information.
Sequence Identity and Sequence-based Evolutionary Distance

![Graph showing the relationship between sequence identity and PAM distance.](image)
Compared to preceding plot, RMSD is better able to pinpoint relationships between more divergent sequences (RMSD stays relatively small for a longer time as compared to PAM distance). Note that the spread around RMSD is larger.
Sequence Searching and Sequence Alignment

- Sequence searching and sequence alignment are different techniques.

- Sequence searching uses sequence alignment, but sequence alignment works on a pre-defined sequence set.

- Sequence searching tries to extract homologous sequences, sequence alignment tries to correctly identify similarities between homologues.
Models Underlying Sequence Searching

Sequence comparison is predominantly based on three models:

1. PAM evolution model
Information content of matched amino acid pair is defined as
\[ \text{PAM score}(XZ) = \log \left[ \frac{P(XZ)}{P(X)P(Z)} \right] \]
with \( P(XZ) \): observed probability of matched XY in trusted alignments
and \( P(X), P(Z) \): probabilities of X and Z in a random sequence

2. Alignment model
local or global alignment of homologues
\rightarrow alignment scores of biologically meaningful alignments

3. Random model
local or global alignment of un-related (=random) sequence pairs
\rightarrow alignment scores of random alignments

Note: Alignment scores are sum of PAM scores + gap penalties
Sequence Searching

QUERY

SEARCHING

SEARCH RESULT

DATABANK

True Negative

False Negative

True Positive

False Positive
Sequence Searching

ROC CURVE DEMONSTRATION

Test value>

Test threshold

http://www.anaesthetist.com/mnm/stats/roc/
Sequence Searching

ROC CURVE DEMONSTRATION

Test value>

Test threshold
Statistics and Thresholds

The simplest statistical selection scheme is to accept only hits above a certain score threshold $T$.

The likelihood of random sequences to yield a score greater than $T$ increases linearly with the logarithm of the 'search space' $n \times m$, where $n$ and $m$ are the sequence lengths.

This gives a formula for accepting hits

$$ S > T + \log(m \times n)/\lambda $$

The parameter $\lambda$ is a scaling constant that is depending upon the scoring scheme (substitution matrix, gap penalties).

>30% sequence identity: homology can be directly inferred

15% - 30% sequence identity: twilight zone

~50% alignments of homologous domains have insignificant scores