

DNA and RNA

1. DNA and RNA are polymer chains composed of chemical compounds of similar structure.
2. The individual units are called nucleotides which are made of three distinct parts: a cyclic base, a cyclic sugar, and a phosphate group (see p. 162, Fig 7-1).
3. There are four different bases, A (adenine), T (thymine), G (guanine), and C (cytosine) in DNA. The RNA code substitutes U (uracil) for T.
4. A and T pair with two hydrogen bonds while G and C pair with three hydrogen bonds. (See p 163, Fig 7-3).
5. DNA has polarity and reads from 5' end to 3' end (see p. 162, Fig 7-1).
6. DNA is of double strands in nature and the two strand are antiparallel.

Gene Finding and DNA Translation

1. Predicting gene location and structure using computational tools.
 - (a) GRAIL uses a neural network to combine evidence from seven different statistical measures of DNA content (Oak Ridge National Lab).
 - (b) GENSCAN utilizes probabilistic models on gene structures. (We will study a bit in detail later.)
 - (c) PROCRUSTES compares potential ORFs with close homologs.
 - (d) GeneWise compares the gene against a single sequence or a model of protein family.
 - (e) Glimmer uses an interpolated Markov model.

Sequence Comparison for Feature Detection

1. A DNA sequence is translated into six seqs (six frame translation), and each sequence is compared to seqs in the database.
2. Each strand requires three frame translation since translation can start at each of the three positions. Since there are two strands, six frame translation is required.
3. Software tools can automatically translate a DNA seq into six frames and compare each of six translated seqs to seqs in the database. **Blastx**.
4. Why we have to bother six translations? Why not compare a DNA seq directly to DNA seq database?

Pairwise Sequence Comparisons

There are major issues in aligning a pair of sequences.

1. How two DNA or AA characters match?

We use scoring matrix that shows how two characters match.

2. How to deal with gaps in the alignment of two sequences?

We penalize gaps. An opening gap and affine gaps are penalized differently.

Why two different gap penalty scores?

3. For example, BLOSUM62 and PAM250 are most widely used matrices.

4. A scoring matrix and the gap penalties are called *a scoring scheme*.

5. Given a scoring scheme, we can compute a score for an alignment of two sequences.

Scoring Matrices

An entry (R_i, C_j) in a scoring matrix has a number, which represents a log-odd ratio of a empirically driven probability that R_i and C_j occur in related sequences to a probability that R_i and C_j occur in randomly occurring sequences.

A *positive value* in an entry (R_i, C_j) means that the two characters, R_i and C_j , are related more than in randomly occurring sequences.

A *negative value* in an entry (R_i, C_j) means that the two characters, R_i and C_j , are related not more than in randomly occurring sequences.

BLOSUM62

Cluster Percentage: >= 62

Entropy = 0.6979, Expected = -0.5209

GAP -12 -2

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4

Sequence Homology and Alignment

There are three different terms commonly used for the (pairwise) sequence analysis: sequence identity, sequence similarity, and sequence homology.

```
>>gi|2688449|gb|AAC66897.1| translation elongation facto (693 aa)
  initn: 1768 init1: 810 opt: 2066 Z-score: 2200.8 bits: 417.7 E():
Smith-Waterman score: 2066; 48.256% identity (49.258% ungapped) in
```

```

                10          20          30          40          50
gi|263 MAREFSLEKTRNIGIMAHIDAGKTTTTERILFYTGRIHKI----GETHEGASQMDWMEQE
      : ::::: ::::: ::::: ::::: ::::: : .. ... :: :: :
gi|268 MDYNKLRNIGISAHIDSGKTTLTERILFYCNKIHA IHEVKGKDGVGAT-MDSMELE
                10          20          30          40          50
```

A Biological Concept and Its Counterpart

In general, a concept in biology is *not* computable and the biological data on the computer cannot be utilized unless there is a computable counterpart.

For example, *homology* is a biological concept that describe the evolutionary relationship among seqs. For the sequence analysis, we use *similarity* and *identity* to model seq homology. It is very very difficult to devise a computable counterpart to a biological concept. Thus seq similarity or identity sometimes do not correspond to seq homology.

Bioinformatics, in a sense, is to devise computable (or searchable) counterpart to biological concepts. The next slide shows how we measure the seq relationship that hopefully model the seq homology.

Dynamic Programming Algorithms for Computing Seq Alignments.

How can we align a pair of seqs?

1. There are numerous possible alignments for a pair of *unaligned* seqs.
2. Given a scoring scheme and an alignment, we can compute a score for the alignment.
3. A naive solution is to enumerate all possible alignments of the two seqs and select the *best scoring* alignment as the alignment of the two scores.
4. However, the number of alignments, especially for *local alignments*, grows exponentially in relation to the length of the seqs. Thus it is not possible to enumerate all possible alignments in search of the best scoring one.
5. The dynamic programming technique is used to reduce computation time.

A Sample Pairwise Sequence Alignment

	-	A	I	H	E	V	K	G	K
-	0	0	0	0	0	0	0	0	0
K	0								
I	0								
G	0								
E	0								

BLAST (Basic Local Alignment Search Tool)

BLAST is the most widely used search tool and it is available on our machine.

Four steps to use BLAST are:

1. Edit `.ncbirc` file to have

```
[NCBI]
```

```
Data=/usr/local/bio/blast-data
```

2. Format the database file

```
formatdb -i yourfile.faa -p T -o T
```

3. Run Blast program, `blastp` for proteins

```
blastall -p blastp -d yourfile.faa -i query.faa -o myoutput
```

4. Look at the search result in `myoutput`

BLAST Search tools

1. `blastall -p program-name`

`blastp, blastn, blastx, tblastn, tblastx`

2. PSI-BLAST, PHI-BLAST

`blastpgp -i query -B alignment -j 2 -d db`

3. `bl2seq -i seq1 -j seq2 -p program-name -g T -o myoutput`

FASTA

1. `fasta33` query database > output
2. `ssearch33` query database > output
3. `fastx33/fasty33` compares a DNA sequence to a protein db.
4. `tfastx33/tfasty33` compares a protein sequence to a DNA db.

Parallel Blast on Solar

See <http://www.indiana.edu/~rac/bioinformatics/solarblast.html>.

You have to create a file `.ncbirc` with the following two lines.

```
[NCBI]
```

```
Data=/miscapps/ncbitoolkit/data
```

An example run of Blast on Solar:

```
$ blastjob -p blastp -i pep.new -d nr -o pep.nr -a 4
```

Interpretation of the Search Result

There are two different kinds of scores, absolute scores (raw score) and statistically adjusted scores (bit score and Zscore).

There are values that are not scores but useful in interpreting the search results. For example, the expectation value (evalue) indicates the number of alignments one expects to find with a score greater than or equal to the observed alignment's score in a search against a *random database*.