# Protein and DNA Sequence Comparison

1. Recent explosion in DNA sequence information = how to interpret this wealth of information

2. Development of computationally efficient methods for detecting sequence similarities

*Useful web sites:*

http://www.fugu-sg.org (genomic databases)
http://www.ncbi.nlm.nih.gov (pointers on databases + NCBI-Blast)
http://www.nature.com/omics/index.html

**Genome sequencing projects statistics**

| Organism | Complete | Draft assembly | In progress | total |
|---|---|---|---|---|
| Prokaryotes | 1117 | 966 | 595 | 2678 |
| Archaea | 100 | 5 | 48 | 153 |
| Bacteria | 1017 | 961 | 547 | 2525 |
| Eukaryotes | 36 | 319 | 294 | 649 |
| Animals | 6 | 137 | 106 | 249 |
| Mammals | 3 | 41 | 25 | 69 |
| Birds | | 3 | 13 | 16 |
| Fishes | | 16 | 16 | 32 |
| Insects | 2 | 38 | 17 | 57 |
| Flatworms | | 3 | 3 | 6 |
| Roundworms | 1 | 16 | 11 | 28 |
| Amphibians | | 1 | | 1 |
| Reptiles | | 2 | | 2 |
| Other animals | | 20 | 24 | 44 |
| Plants | 5 | 33 | 80 | 118 |
| Land plants | 3 | 29 | 73 | 105 |
| Green Algae | 2 | 4 | 6 | 12 |
| Fungi | 17 | 107 | 59 | 183 |
| Ascomycetes | 13 | 83 | 38 | 134 |
| Basidiomycetes | 2 | 16 | 11 | 29 |
| Other fungi | 2 | 8 | 10 | 20 |
| Protists | 8 | 39 | 46 | 93 |
| Apicomplexans | 3 | 11 | 16 | 30 |
| Kinetoplasts | 4 | 3 | 2 | 9 |
| Other protists | 1 | 24 | 28 | 53 |
| **total:** | **1153** | **1285** | **889** | **3327** |

Revised: Feb 16, 2012

http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html

# Which organisms have been sequenced?

Full genome sequencing—many other (less involved/costly) methods exist

-What do you want to learn from sequencing?

Humans—James Watson, Craig Venter, Yang Huanming, Seong-Jin Kim

Model organisms—E. coli, Fission and budding yeast species, Drosophila melanogaster, C. elegans, Mus musculus (lab mouse), Danio rerio (zebrafish), Zea mays (corn/maize), Arabidopsis thaliana

Other organisms—chicken, puffer fish, cow, dog, guinea pig, cat, elephant, rhesus, mouse, bat, mosquito, honey bee, many prokaryotes (mostly bacteria)

Metagenomics—Sequencing many organisms at once (without separation by species), with the intent of understanding complex microbial ecologies
Human gut microbes, deep sea vents, ocean, soil…
http://www.nature.com/nrmicro/focus/metagenomics/index.html

# So much data—how do we view?

Primary sequence (CGAT…) is not very informative without annotation. A big part of genome sequencing is generating and presenting annotation of the primary DNA sequence.

Lots of work done on this—how to access?

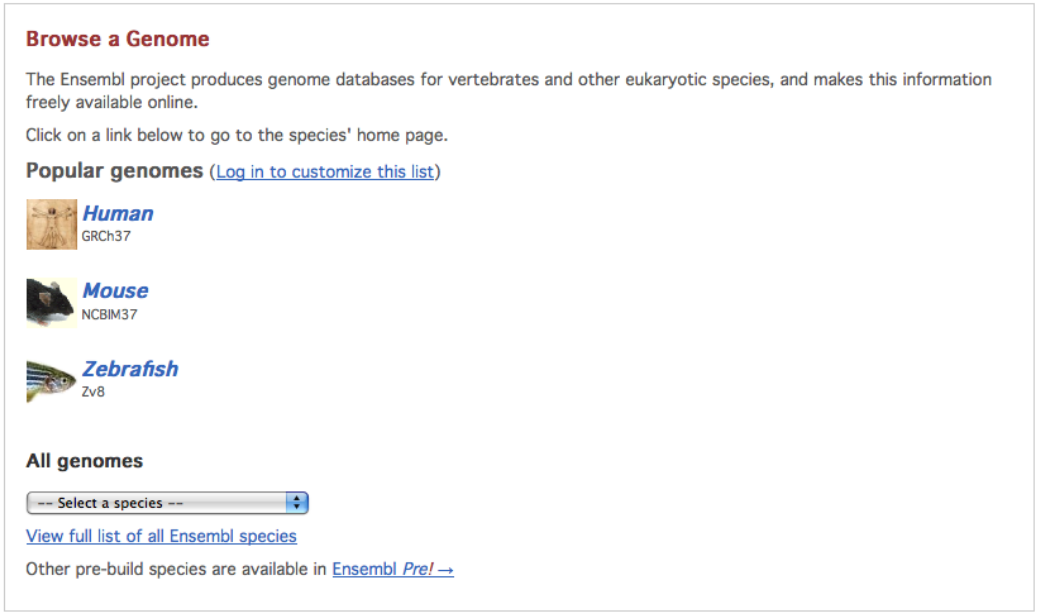
Gateways:

http://www.ncbi.nlm.nih.gov/guide/genomes-maps/

http://www.ebi.ac.uk/genomes/

http://uswest.ensembl.org/index.html

Lots to explore here!



# Sequence Comparison: why and how

1. **Automated methods for comparing DNA or protein sequences:**
   - most common and most powerful method for protein structure/function prediction
   - responsible for much of the rapid progress in biology over last 5-15 years (realization that similar processes underlie the development of most organisms, etc)
   - interesting parallels to protein folding problem

2. **Two components:**

- a scoring matrix: evaluate an alignment (identity works well for DNA, for amino-acids it is better to give non-zero scores for conservative mutations)

- an alignment algorithm: given the scoring matrix, find the best alignment possible

| | A | T | C | G |
|---|---|---|---|---|
| **A** | 1 | 0 | 0 | 0 |
| **T** | 0 | 1 | 0 | 0 |
| **C** | 0 | 0 | 1 | 0 |
| **G** | 0 | 0 | 0 | 1 |

Identity scoring matrix for DNA

# Part 1: Scoring Matrixes

Scoring matrices: Dayhoff's and Henikoff's

- Dayhoff aligned many pairs of sequences with more than 85% sequence identity and evaluated the frequencies of occurence of all amino acid pairs
- the expected frequency of substitutions in more distantly related pairs was obtained by extrapolation (multiply substitution matrix by itself many times)
- want to know whether alignment is more likely than one between unrelated sequences => divide by probability of substitution occuring by chance
  - log-odds matrix $\log (p_{ij}/p_i p_j)$
- Henikoff generated an improved matrix, BLOSSUM62, by directly evaluating substitution frequencies in multiple sequence alignments for protein families rather than extrapolating from pairs of closely related sequences.

Dayhoff *et al.*, A model of evolutionary change in proteins (1978) *in* "Atlas of Protein Sequence and Structure" 5(3) M.O. Dayhoff (ed.), 345 - 352, National Biomedical Research Foundation, Washington

Henikoff, S and Henikoff, J.,G., *Amino acid substitution matrices from protein blocks*,1992 PNAS **(89),** 10915 - 10919.

# The BLOSSUM62 substitution table:

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

- typical gap penalties used with this table are -11 for opening a gap and -1 for each residue in the gap.
- which is a better alignment??

  A I K OR A I K OR A I K

  V V A I A V A _ K K

# Part 2: The Alignment Problem

- given a scoring matrix, how to find optimal alignment?
- need to allow for gaps and insertions (evolution)
- huge combinatorics problem:

  sequence 1: atcgctaatgcctagccatttgcaagac

  sequence 2: tcaagtccaatgccgaaattgcaagtac
- for two sequences 300 residues long, ~ $10^{88}$ alignments (can't try all of them!)
- elegant solution: dynamic programming algorithm

  IF

  #1 A G T G C A

  #2 A G - G C T

  is an optimal alignment,

  THEN

  #1 A G T G C

  #2 A G - G C

  must also be optimal, etc

  (if not, could improve overall alignment by altering subalignments)

Example: Align AGGC with AATGC using identity matrix and no gap penalty.

| | | A | A | T | G | C |
|---|---|---|---|---|---|---|
| | | | | | | |
| A | | | | | | |
| A | | | | | | |
| G | | | | | | |
| C | | | | | | |

- Each entry = score for aligning pair of residues with optimal alignment of previous residues
- Dynamic programming algorithm
  - requires time ~ length$^2$ rather than ~length$^{(length)}$
  - works because interactions are **local:**
    - score for whole = sum of scores for parts (cf protein folding)
- BLAST, FASTA  more efficient approximate solutions to alignment problem

# Estimating how a good an alignment is

To assess whether a given alignment constitutes evidence for homology, need to know how strong an alignment can be expected from chance alone.

Assessment of alignment significance is also critical to the iterative methods discussed in a few slides.

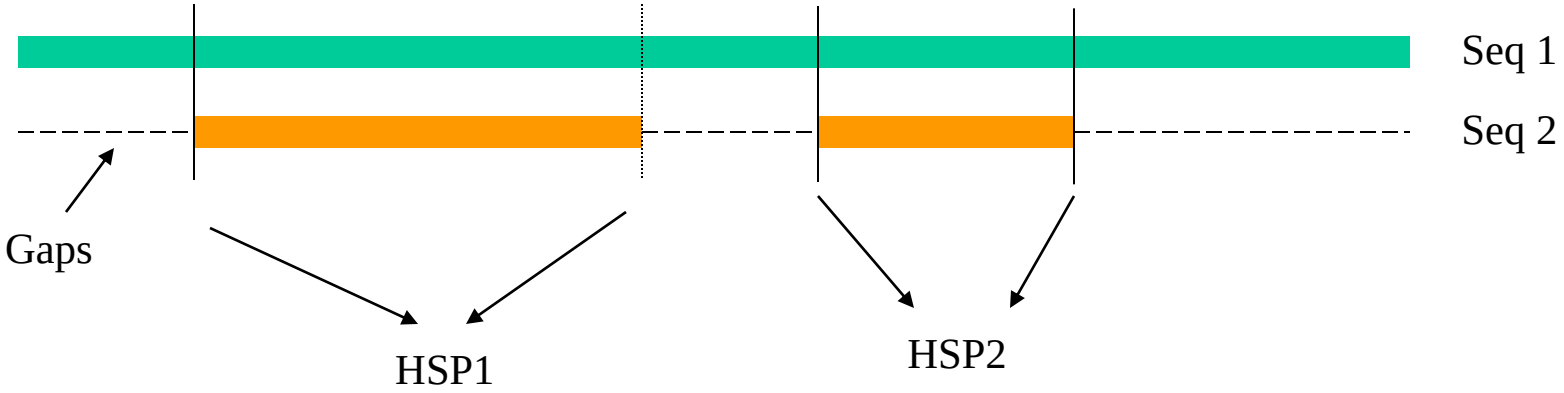
**How ?**

*E value* : number of matches expected with score > S
*P value* : probability of finding a match with score > S
(The two are related: P = 1 - exp (-E) )

How reliable is a match with an E value of 1.0 ?
of .00001 ?

# How are E-values computed ?

A local alignment without gaps consists simply of a pair of equal length segments, one from each of the two sequences being compared, whose scores can not be improved by extension or trimming. These are called high-scoring segment pairs or HSPs.




To analyze how high a score is likely to arise by chance, a model of random sequences is needed. For proteins, the simplest model chooses the amino acid residues in a sequence independently, with specific background probabilities for the various residues.

# An analytical expression for the E-value

In the limit of sufficiently large sequence lengths *m* and *n*, the statistics of HSP scores are characterized by two parameters, *K* andω. The expected number of HSPs with score at least *S (the E value)* is given by the formula

$$\mathbf{E = K\, m\, n \exp(-\omega\, S)}$$

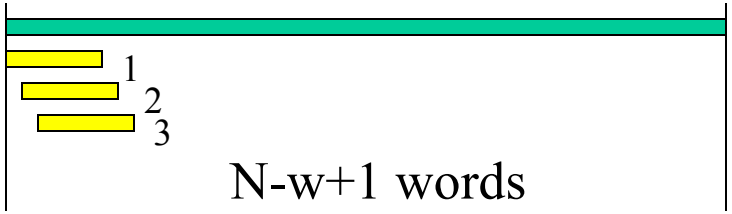
This formula makes intuitive sense. Doubling the length of either sequence should double the number of HSPs attaining a given score. Also, for an HSP to attain the score *2x* it must attain the score *x* twice in a row, so one expects *E* to decrease exponentially with score. The parameters *K* and ω can be thought of simply as natural scales for the search space size and the scoring system respectively.

# BLAST: a faster heuristic algorithm

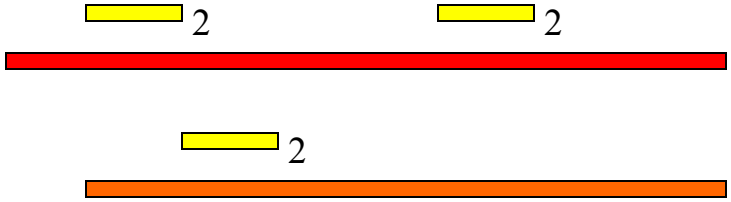
Dynamic programming always finds the best global alignment between 2 sequences of size m and n, but in a time which is proportional to mn.

For searching for a query sequence in a Genomic DB, this is too slow!
BLAST is a different approach that rapidly finds significant local sequence matches between a query sequence and sequences in a database

1) query sequence is divided into words of size w (generally w=11) for comparing DNA sequences

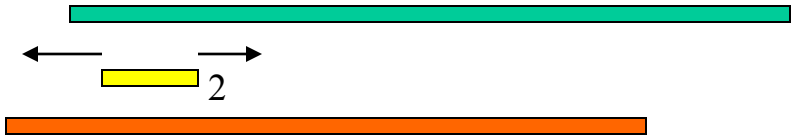



2) Matches are searched for each word in the full database. The score of each match found, S, is compared to a threshold T. If S>T, the match is called a *hit* and kept.




Hits in DB

3) For each hit, the alignment is grown on the left and right till the score stops growing.
This results in a set of HSP's




Extending hits to find HSPs

# BLAST (ctd..)

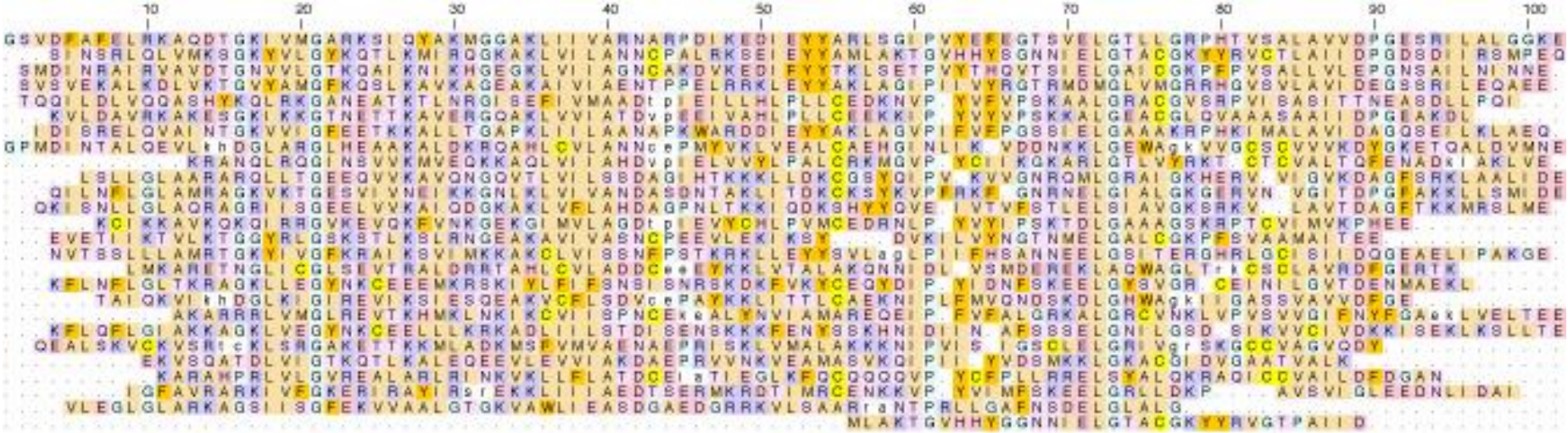
4) total score for each sequence of the database is the sum of the HSPs found for that sequence, if any.

Advantages of BLAST:

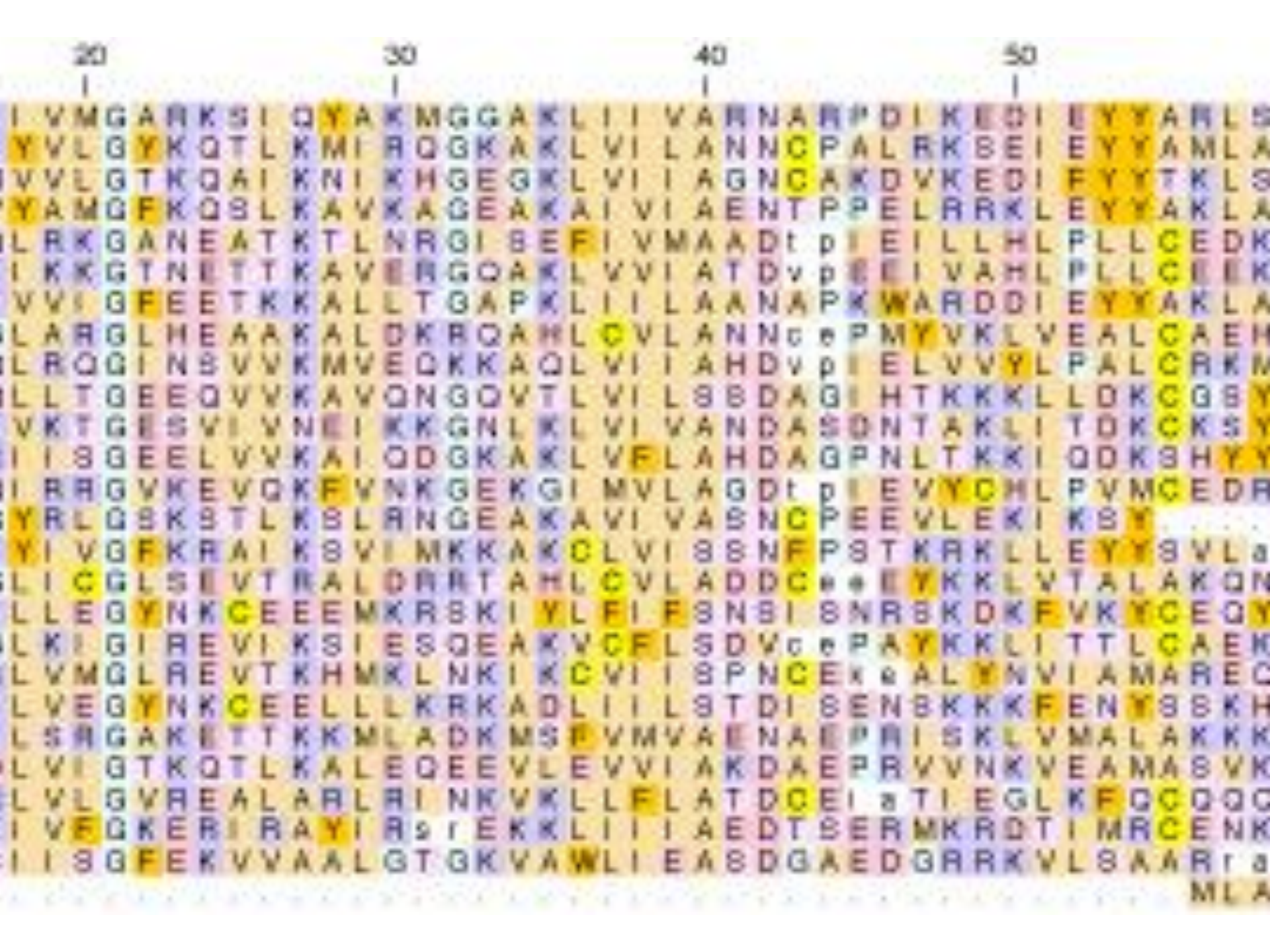
- fast, allows searching of complete databases
- find local alignments that may be biologically significant, but hard to find with other methods
- the search algorithm can be used iteratively: PSI-BLAST

Ref: Altschul, S.,F., *et al.*, Basic Local Alignment Search Tool, *JMB*, 1990, **215**, 403-410

# Improvements to the Method Using Multiple Sequence Alignments



Multiple Sequence Alignments (MSA) contain a wealth of information that can be used to improve sequence searching methods


20
30
40
50

# The Information in the MSA can be used in different ways

1. Improved substitution matrices. BLOSSUM62 (Henikoff)

2. Profile methods:
   - previous methods utilize single substitution matrix at all positions, but at different positions in proteins, different residues are likely to substitute for each other.
   - if you have a number of related sequences, you can obtain family specific substitution frequencies directly from multiple sequence alignment.
   - You can use position specific scoring matrix with dynamic programming algorithm as before.
   - can progressively build up better and better position specific scoring matrix by iteration: search database, add new sequences to multiple sequence alignment, generate new scoring matrix, repeat. This is the basic idea behind PSI-BLAST, probably the best current method.
   - http://www.ncbi.nlm.nih.gov/BLAST/

# The PSI-BLAST Methodology

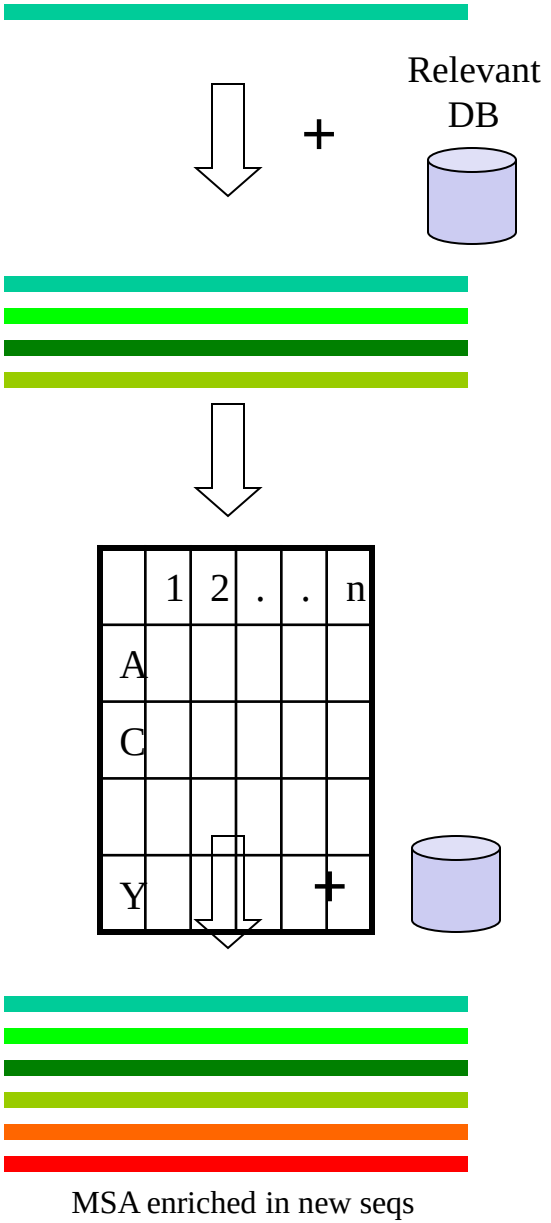
1. PSI-BLAST takes as an input a single protein sequence and compares it to a protein database, using BLAST.

2. The program constructs a multiple alignment, and then a profile, from any local alignments above a specified E value cutoff. Different numbers of sequences can be aligned in different template positions.

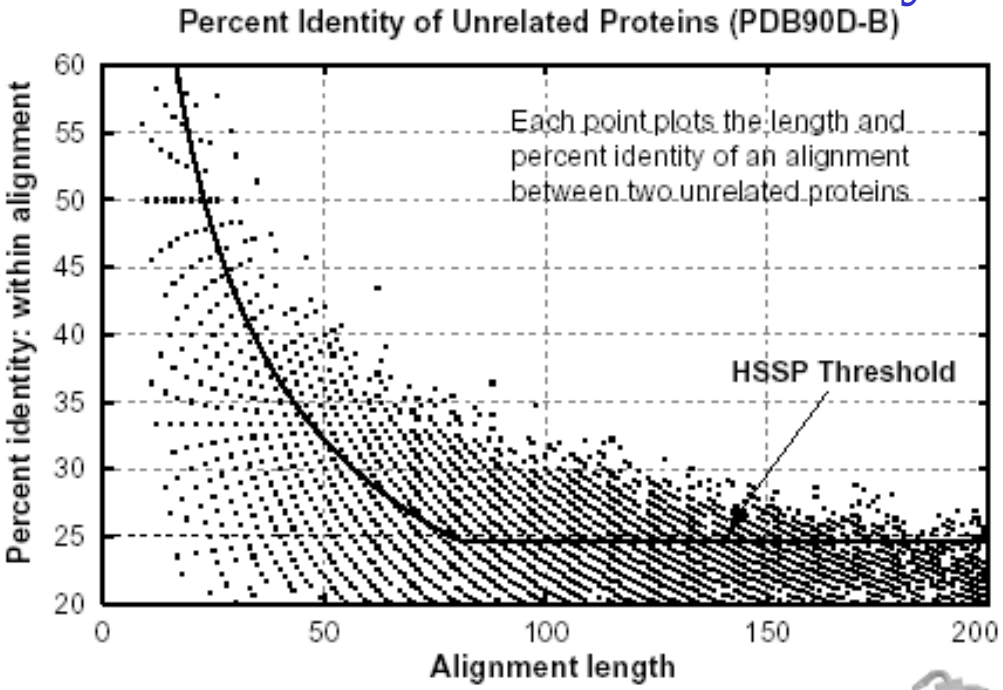
3. The profile is compared to the protein database, again seeking local alignments.

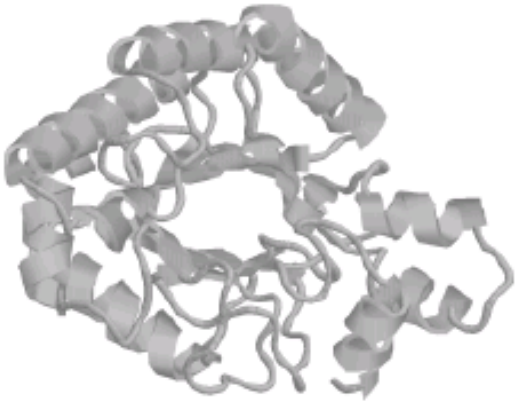
4. PSI-BLAST estimates the E values of all local alignments found. Because profile substitution scores are constructed to a fixed scale, and gap scores remain independent of position, the statistical theory and parameters for BLAST alignments remain applicable to profile alignments.

5. Finally, PSI-BLAST iterates, by returning to step (2), an arbitrary number of times or until convergence



# The relationship between sequence similarity and structural/functional similarity can be assessed empirically



Hemoglobin β-chain (1hdsb)  Cellulase E2 (1tml_)

```
1hdsb GKVDVDVVGAQALGR--LLVVYPWTQRFFQHFGNLSSAGAVMNNPKVKAHGKRVLDAFTQGLKH
      :.::. . .::: :.  .:::: :     :: ::.::    :. .:. .  .: :. :::.
1tml_ GQVDALMSAAQAAGKIPILVVYNAPGR---DCGNHSSGGA----PSHSAY-RSWIDEFAAGLKN
```

# References

Sequence comparisons methods and algorithms are not covered in the reference books. However:

- *Biological Sequence Analysis*, by R.Durbin, S.Eddy, A. Krogh and G. Mitchison (Cambridge Univ. Press) has a thorough coverage of all state-of-the-art algorithm used for sequence analysis (contains dynamic programming as well as other topics like HMM and formal grammars)

- Several monographies exist on BLAST alone: BLAST, by I. Korf, M. Yandell and J. Bedell (O'Reilly eds.) explains the algorithm as well as how to actually use BLAST efficiently for biological research.

End of lecture