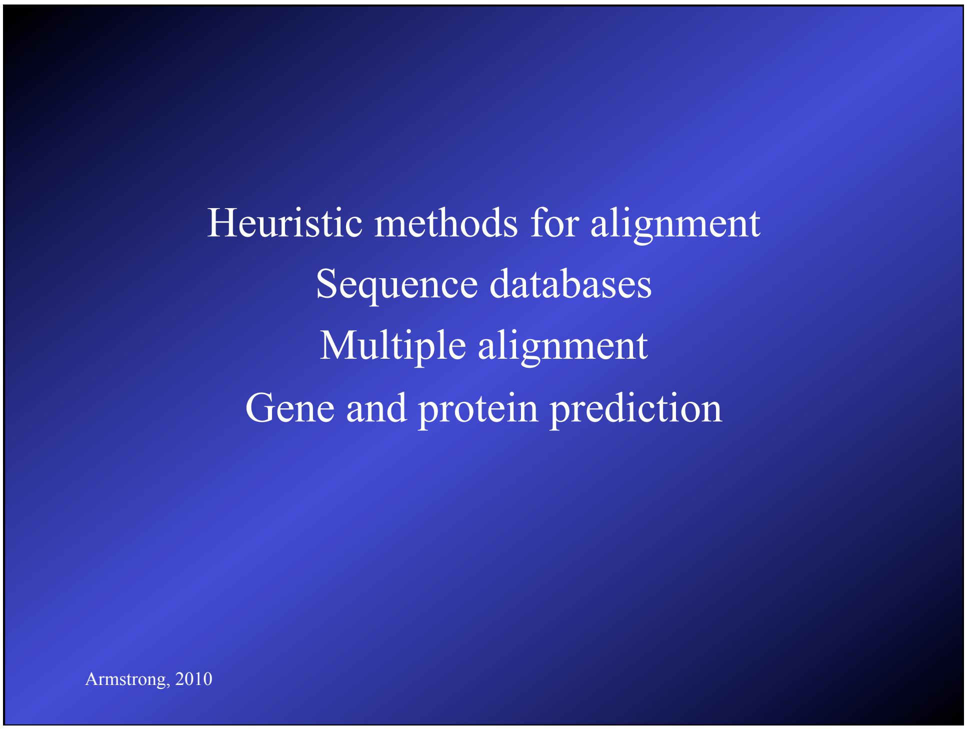
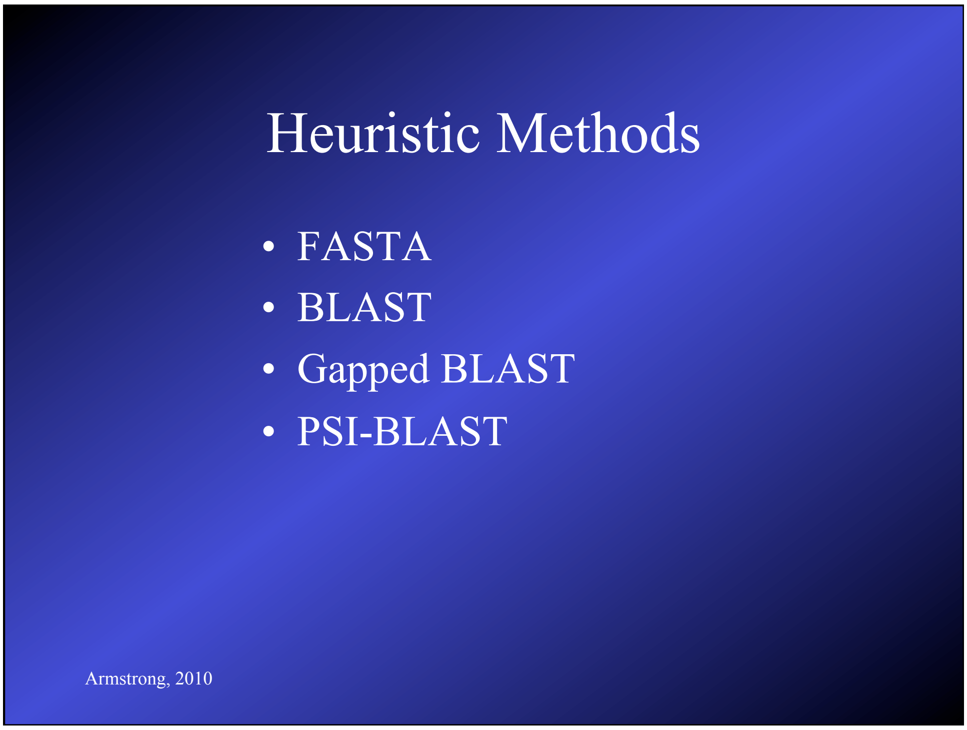


Heuristic methods for alignment
Sequence databases
Multiple alignment
Gene and protein prediction

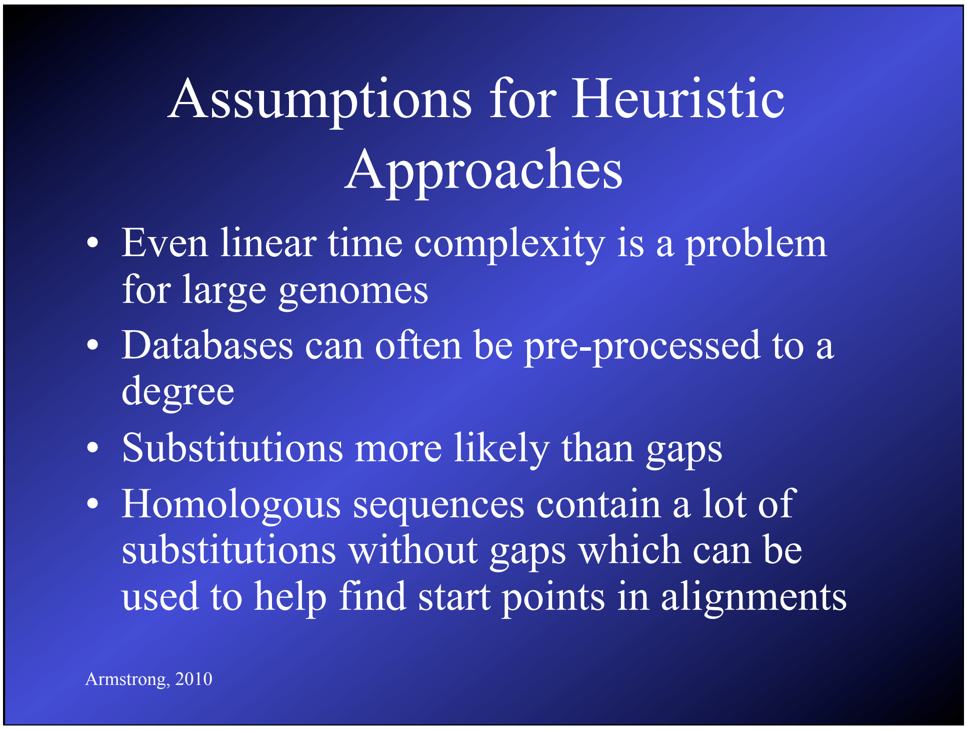
Armstrong, 2010



Heuristic Methods

- FASTA
- BLAST
- Gapped BLAST
- PSI-BLAST

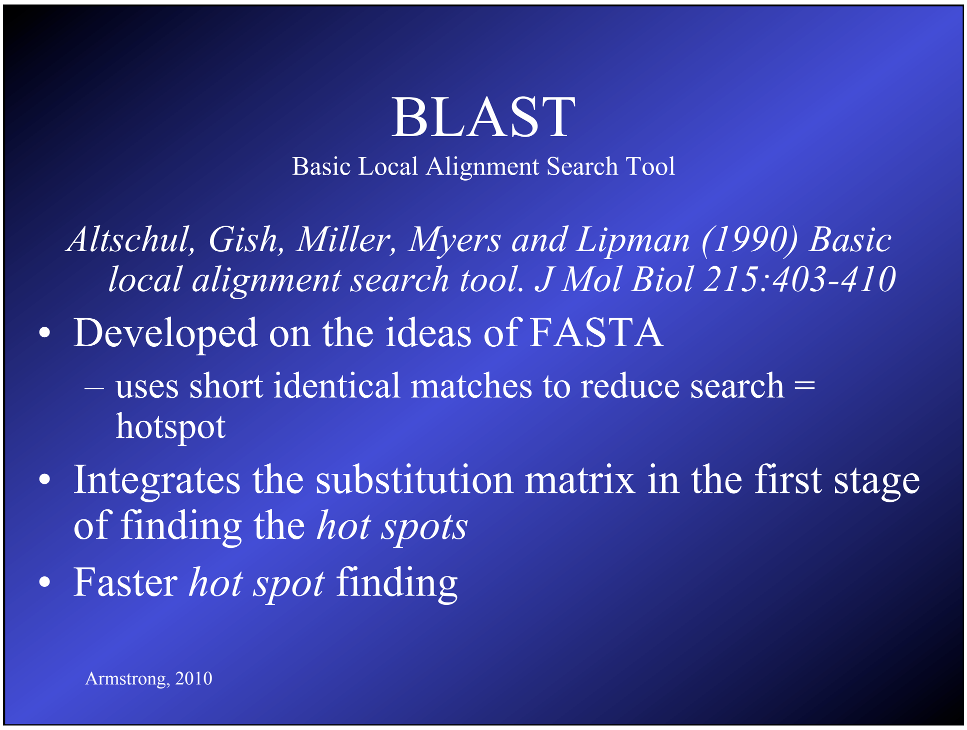
Armstrong, 2010



Assumptions for Heuristic Approaches

- Even linear time complexity is a problem for large genomes
- Databases can often be pre-processed to a degree
- Substitutions more likely than gaps
- Homologous sequences contain a lot of substitutions without gaps which can be used to help find start points in alignments

Armstrong, 2010



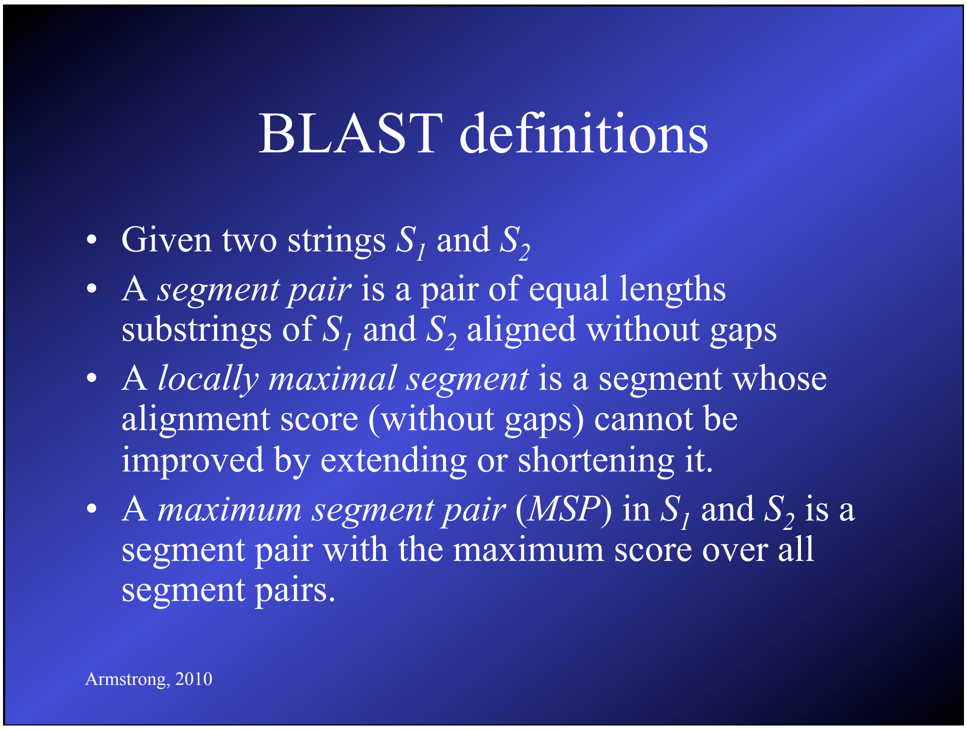
BLAST

Basic Local Alignment Search Tool

Altschul, Gish, Miller, Myers and Lipman (1990) Basic local alignment search tool. J Mol Biol 215:403-410

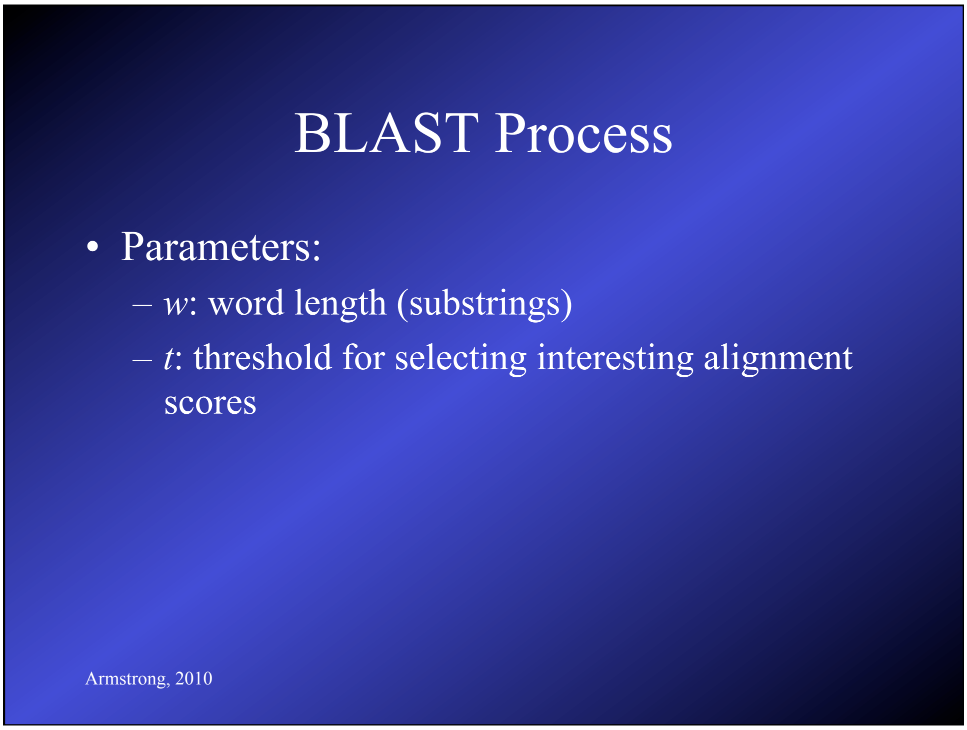
- Developed on the ideas of FASTA
  - uses short identical matches to reduce search = hotspot
- Integrates the substitution matrix in the first stage of finding the *hot spots*
- Faster *hot spot* finding

Armstrong, 2010



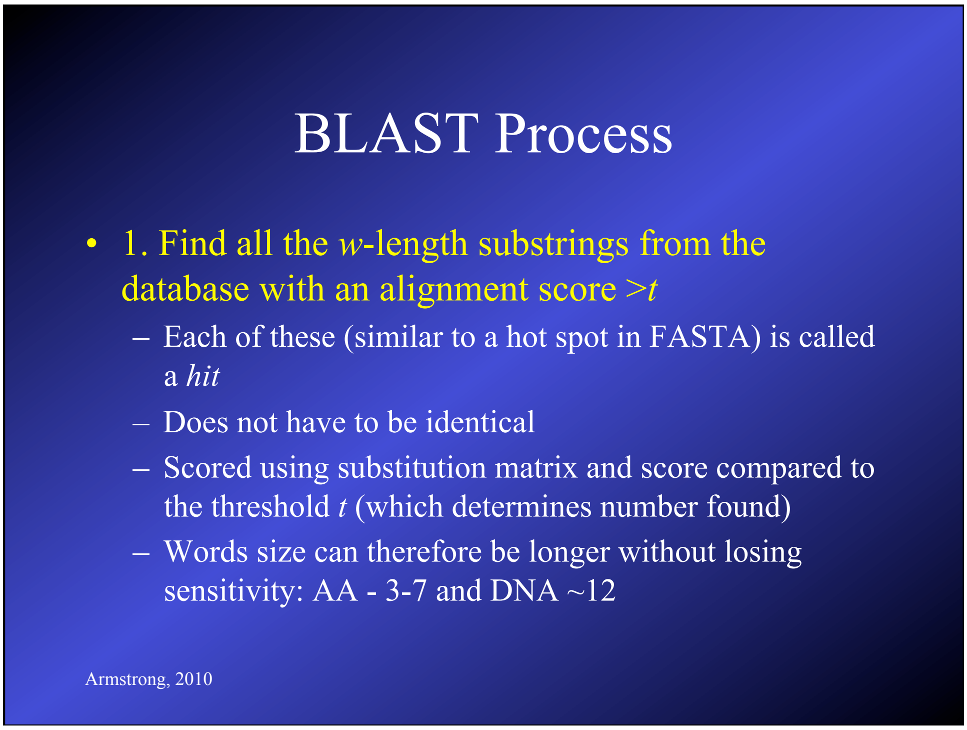
BLAST definitions

- Given two strings S_1 and S_2
- A *segment pair* is a pair of equal lengths substrings of S_1 and S_2 aligned without gaps
- A *locally maximal segment* is a segment whose alignment score (without gaps) cannot be improved by extending or shortening it.
- A *maximum segment pair* (*MSP*) in S_1 and S_2 is a segment pair with the maximum score over all segment pairs.



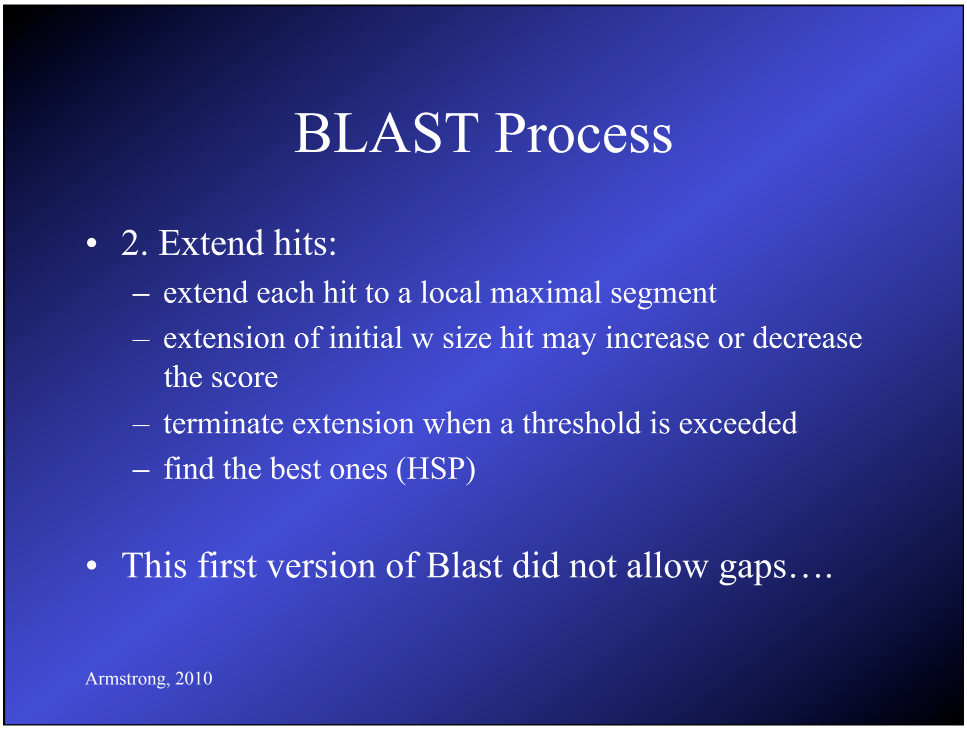
BLAST Process

- Parameters:
  - *w*: word length (substrings)
  - *t*: threshold for selecting interesting alignment scores



BLAST Process

- 1. Find all the *w*-length substrings from the database with an alignment score >*t*
  - Each of these (similar to a hot spot in FASTA) is called a *hit*
  - Does not have to be identical
  - Scored using substitution matrix and score compared to the threshold *t* (which determines number found)
  - Words size can therefore be longer without losing sensitivity: AA - 3-7 and DNA ~12



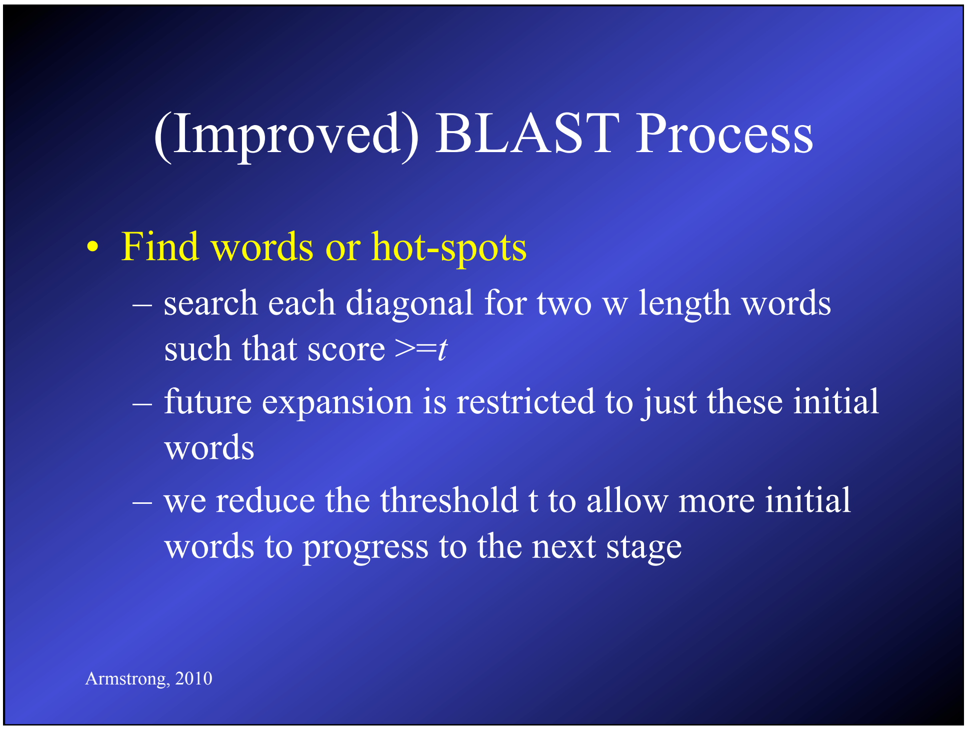
BLAST Process

- 2. Extend hits:
  - extend each hit to a local maximal segment
  - extension of initial w size hit may increase or decrease the score
  - terminate extension when a threshold is exceeded
  - find the best ones (HSP)
- This first version of Blast did not allow gaps….

(Improved) BLAST

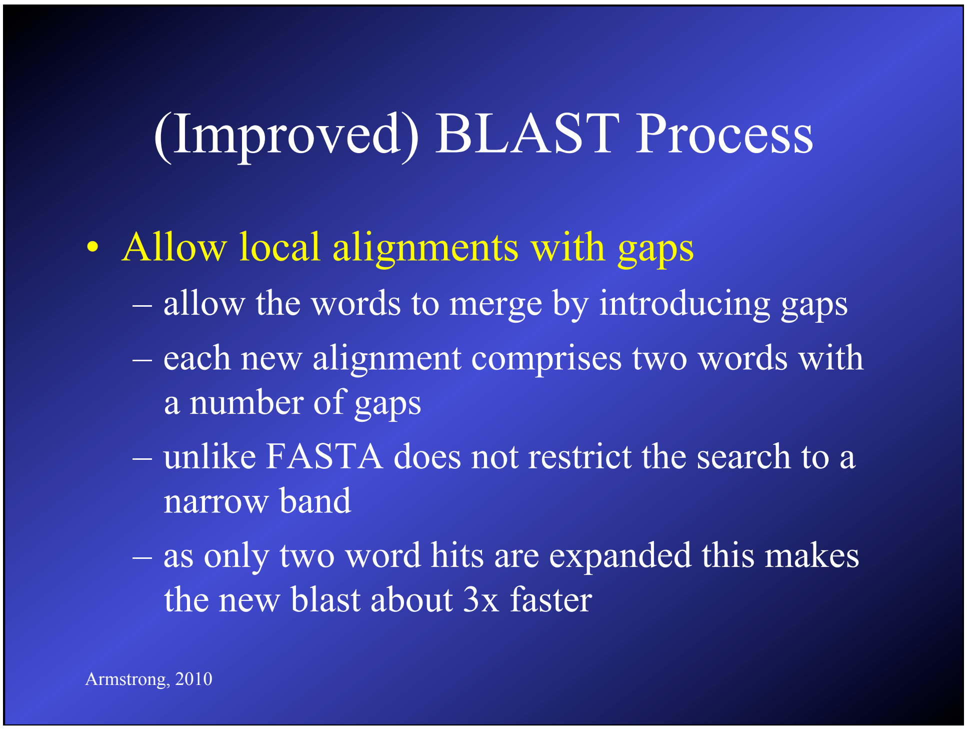
Altshul, Madden, Schaffer, Zhang, Zhang, Miller & Lipman (1997) Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402

- Improved algorithms allowing gaps
  - these have superceded the older version of BLAST
  - two versions: Gapped and PSI BLAST



(Improved) BLAST Process

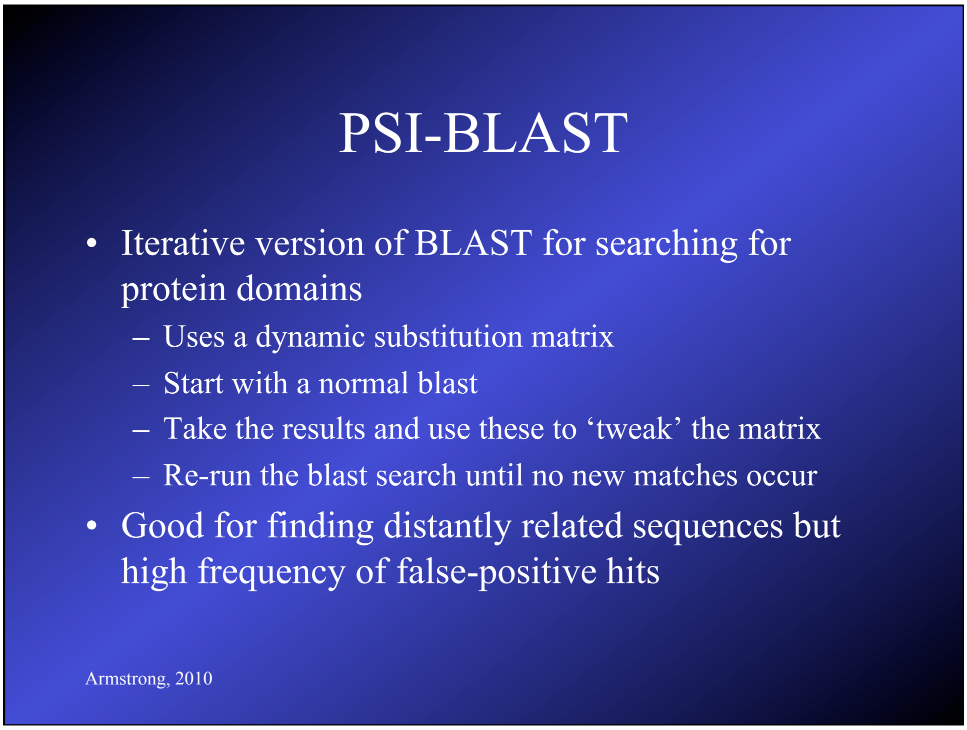
- Find words or hot-spots
  - search each diagonal for two w length words such that score >=*t*
  - future expansion is restricted to just these initial words
  - we reduce the threshold t to allow more initial words to progress to the next stage



(Improved) BLAST Process

- Allow local alignments with gaps
  - allow the words to merge by introducing gaps
  - each new alignment comprises two words with a number of gaps
  - unlike FASTA does not restrict the search to a narrow band
  - as only two word hits are expanded this makes the new blast about 3x faster

Armstrong, 2010



PSI-BLAST

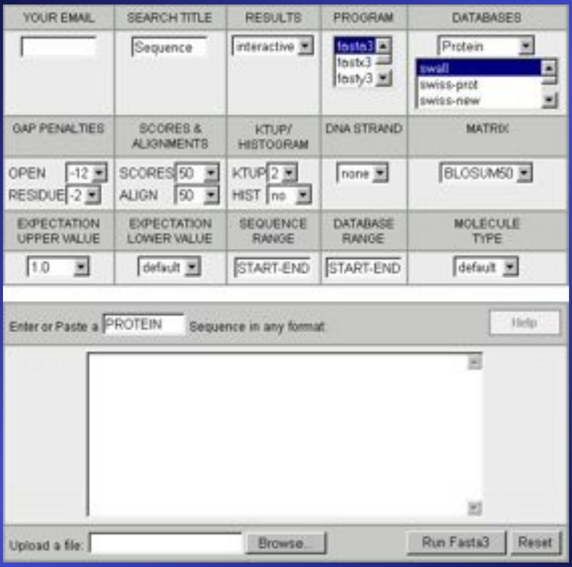
- Iterative version of BLAST for searching for protein domains
  - Uses a dynamic substitution matrix
  - Start with a normal blast
  - Take the results and use these to ‘tweak’ the matrix
  - Re-run the blast search until no new matches occur
- Good for finding distantly related sequences but high frequency of false-positive hits

Armstrong, 2010

BLAST Programs

- blastp compares an amino acid query sequence against a protein sequence database.
- blastn compares a nucleotide query sequence against a nucleotide sequence database.
- blastx compares a nucleotide query sequence translated in all reading frames against a protein sequence database.
- tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. (SLOW)

Armstrong, 2010

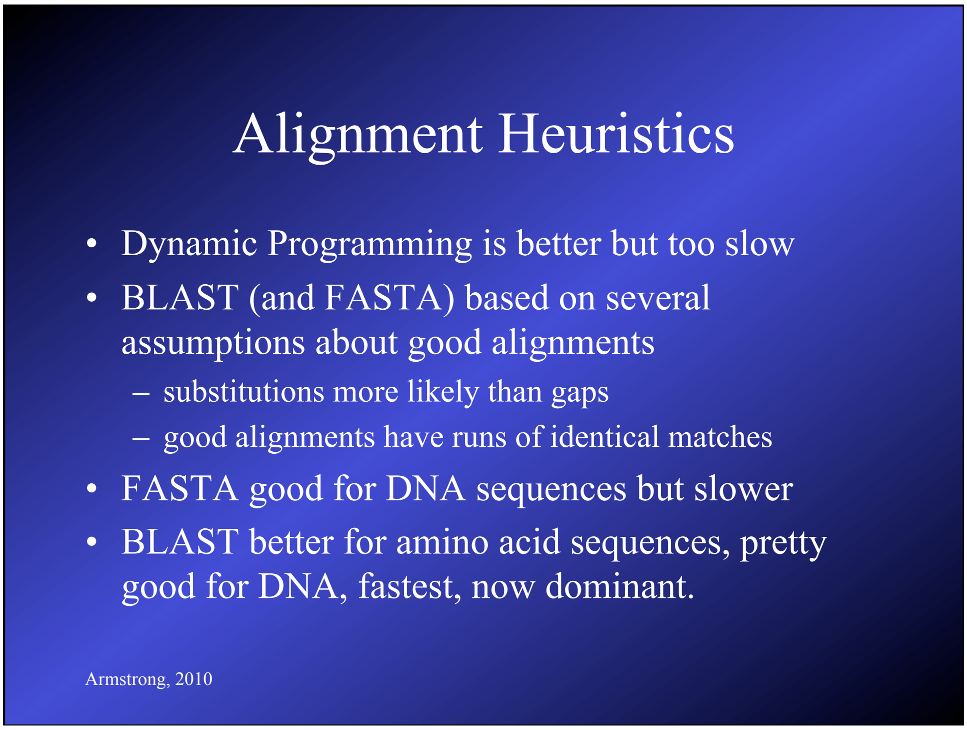


YOUR EMAIL	SEARCH TITLE	RESULTS	PROGRAM	DATABASES
	Sequence	interactive	fasta3 fastx3 fasty3	Protein swall swiss-prot swiss-new
GAP PENALTIES	SCORES & ALIGNMENTS	KTUP/ HISTOGRAM	DNA STRAND	MATRIX
OPEN -12 RESIDUE -2	SCORES 50 ALIGN 50	KTUP 2 HIST no	none	BLOSUM50
EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE	SEQUENCE RANGE	DATABASE RANGE	MOLECULE TYPE
1.0	default	START-END	START-END	default

Enter or Paste a PROTEIN Sequence in any format

Upload a file:

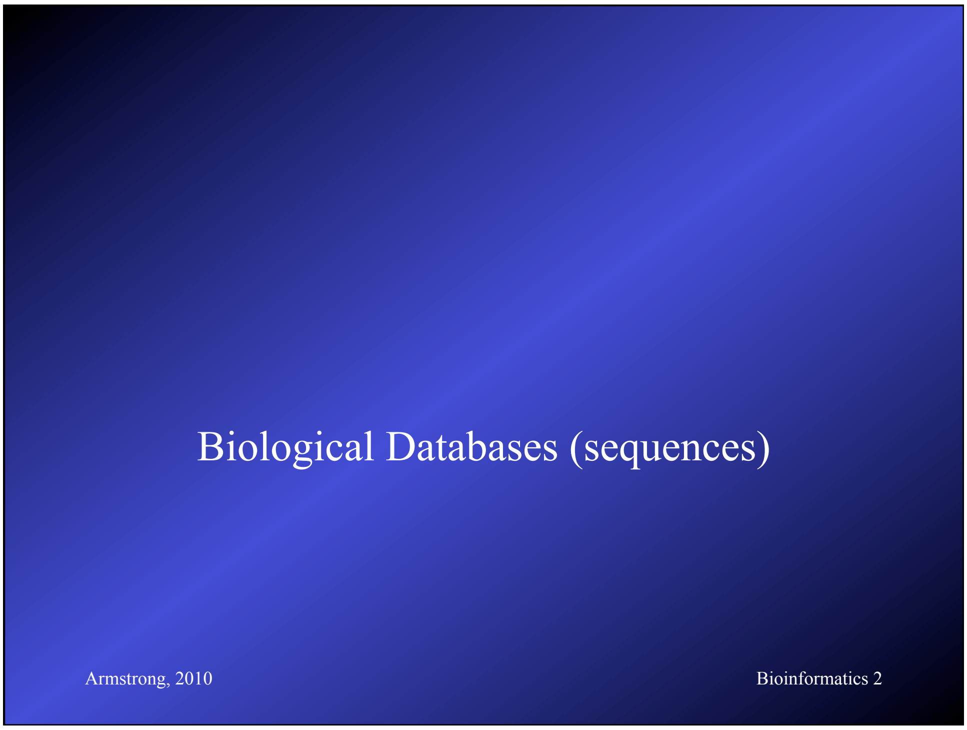
Armstrong, 2010



Alignment Heuristics

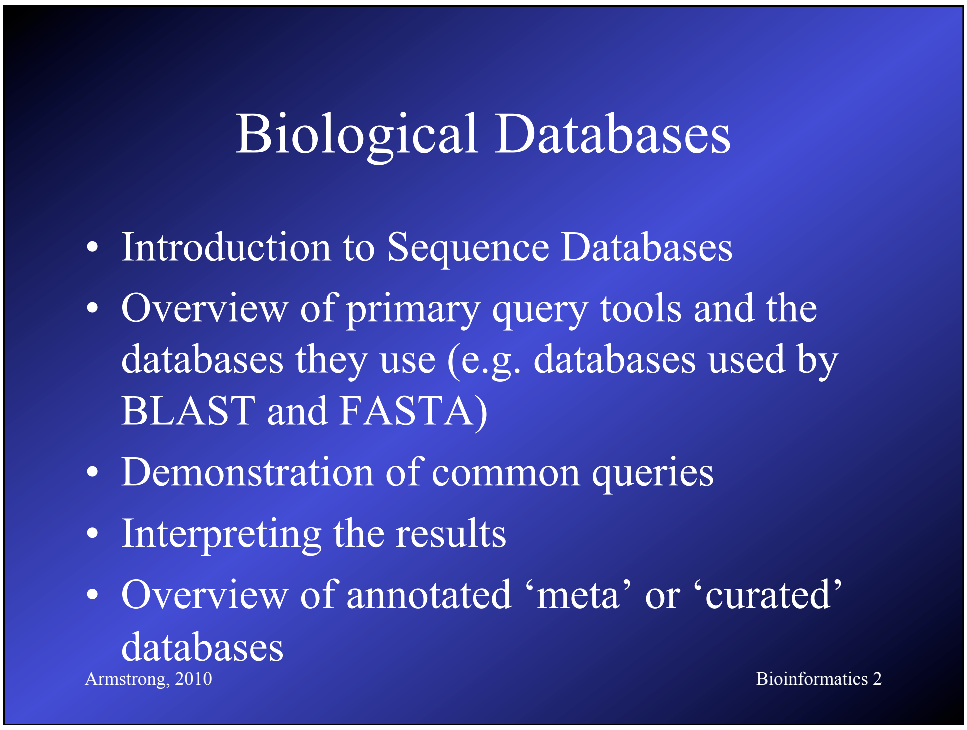
- Dynamic Programming is better but too slow
- BLAST (and FASTA) based on several assumptions about good alignments
  - substitutions more likely than gaps
  - good alignments have runs of identical matches
- FASTA good for DNA sequences but slower
- BLAST better for amino acid sequences, pretty good for DNA, fastest, now dominant.

Armstrong, 2010



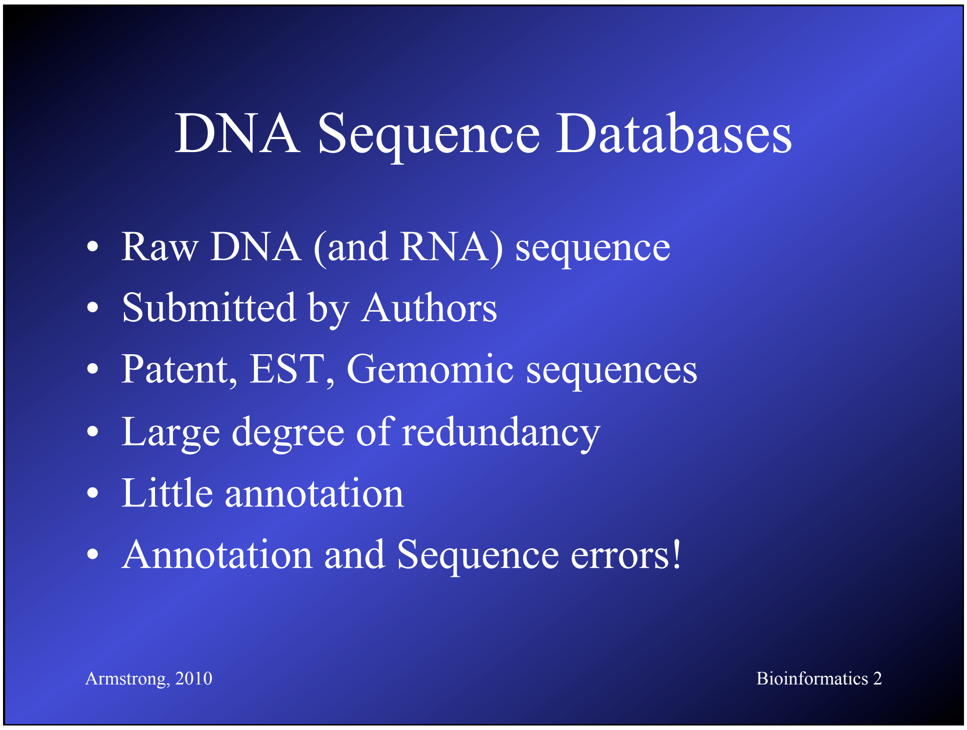
Biological Databases (sequences)

Armstrong, 2010 Bioinformatics 2



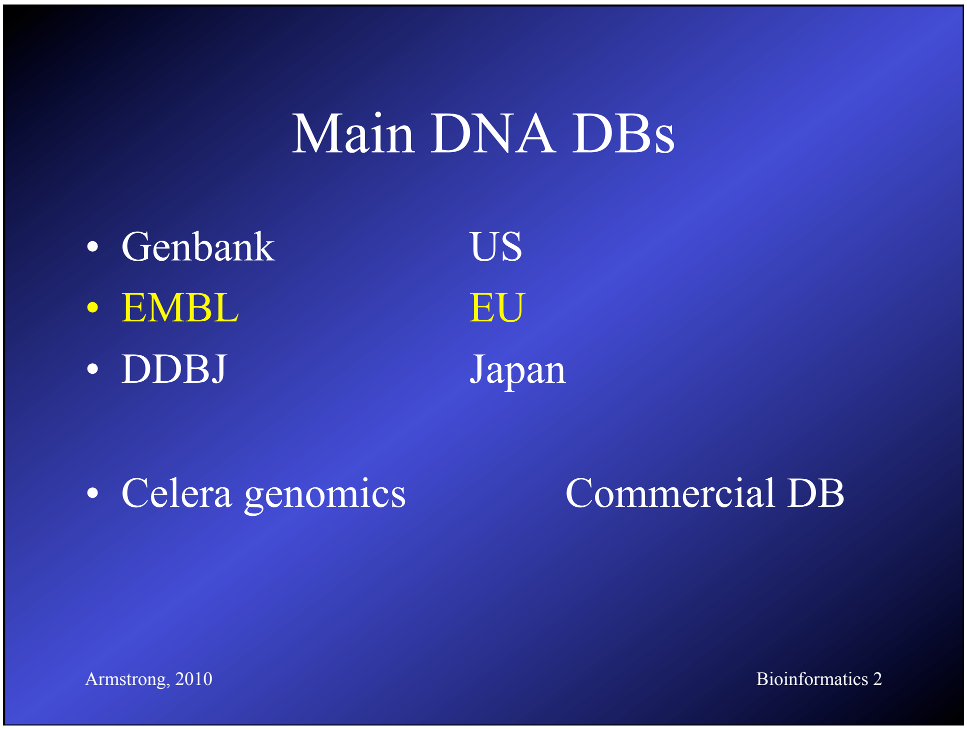
Biological Databases

- Introduction to Sequence Databases
- Overview of primary query tools and the databases they use (e.g. databases used by BLAST and FASTA)
- Demonstration of common queries
- Interpreting the results
- Overview of annotated ‘meta’ or ‘curated’ databases



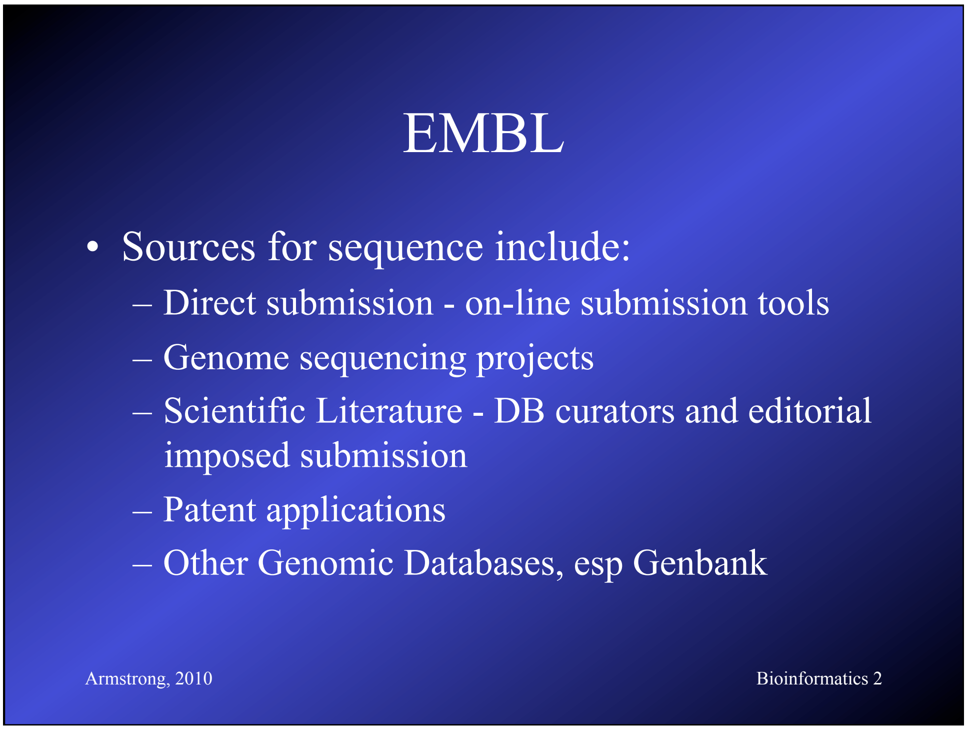
DNA Sequence Databases

- Raw DNA (and RNA) sequence
- Submitted by Authors
- Patent, EST, Gemomic sequences
- Large degree of redundancy
- Little annotation
- Annotation and Sequence errors!



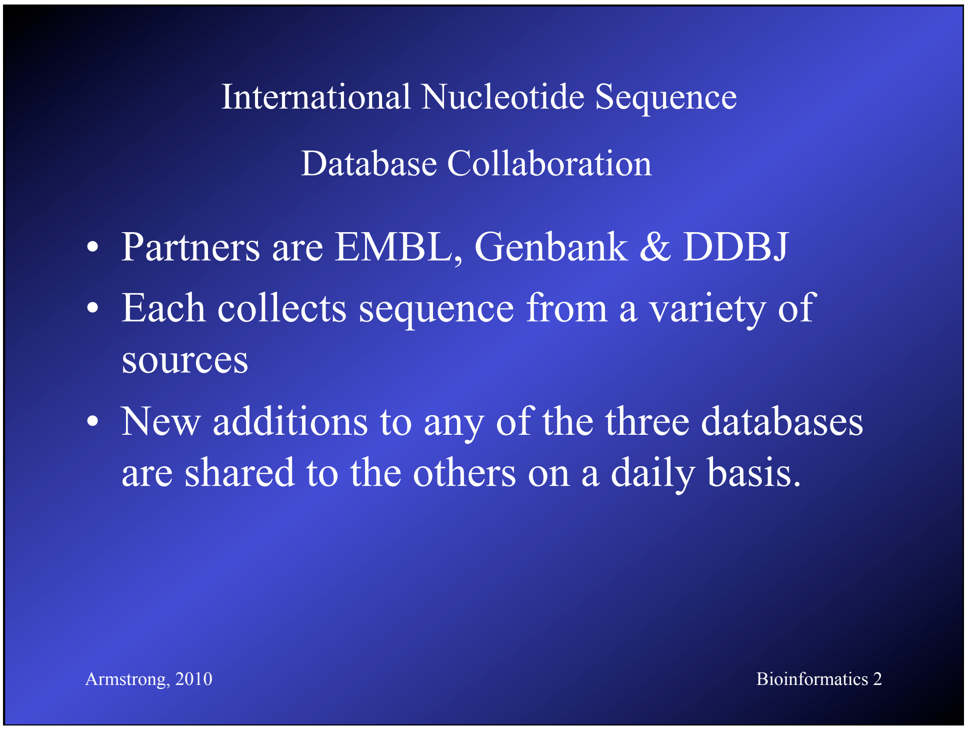
Main DNA DBs

• Genbank	US
• EMBL	EU
• DDBJ	Japan
• Celera genomics	Commercial DB



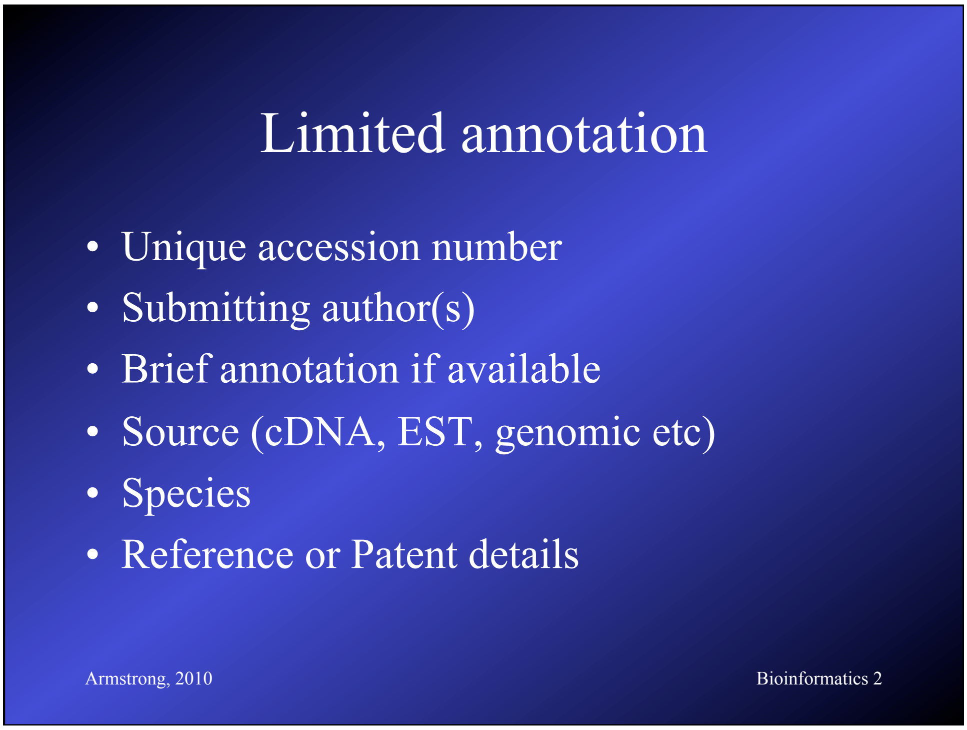
EMBL

- Sources for sequence include:
  - – Direct submission - on-line submission tools
  - – Genome sequencing projects
  - – Scientific Literature - DB curators and editorial imposed submission
  - – Patent applications
  - – Other Genomic Databases, esp Genbank



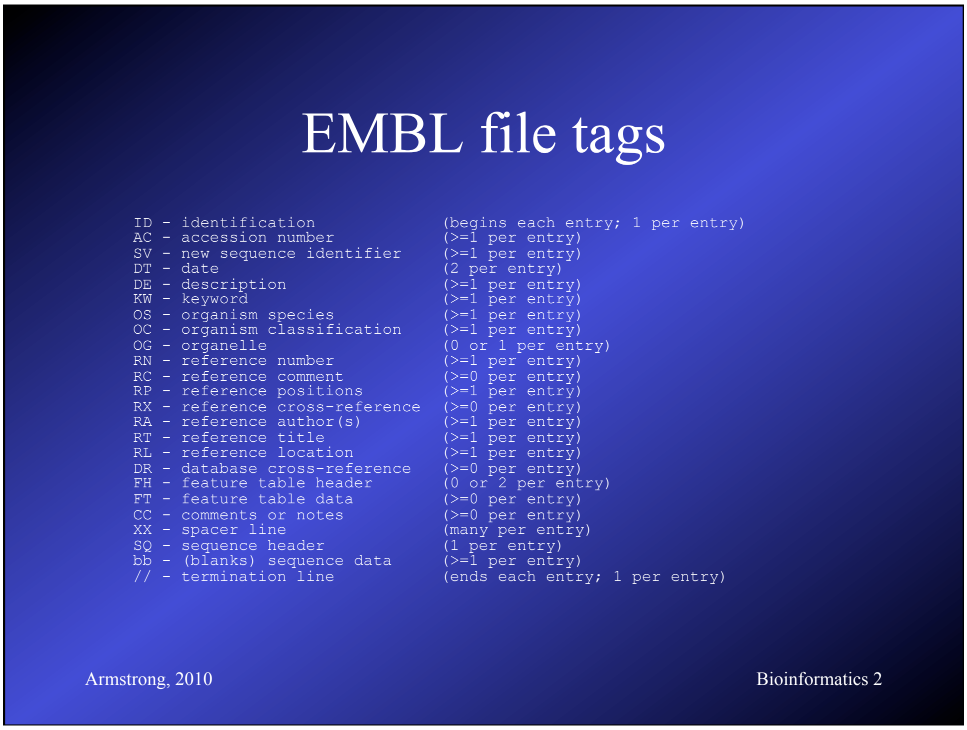
International Nucleotide Sequence Database Collaboration

- Partners are EMBL, Genbank & DDBJ
- Each collects sequence from a variety of sources
- New additions to any of the three databases are shared to the others on a daily basis.



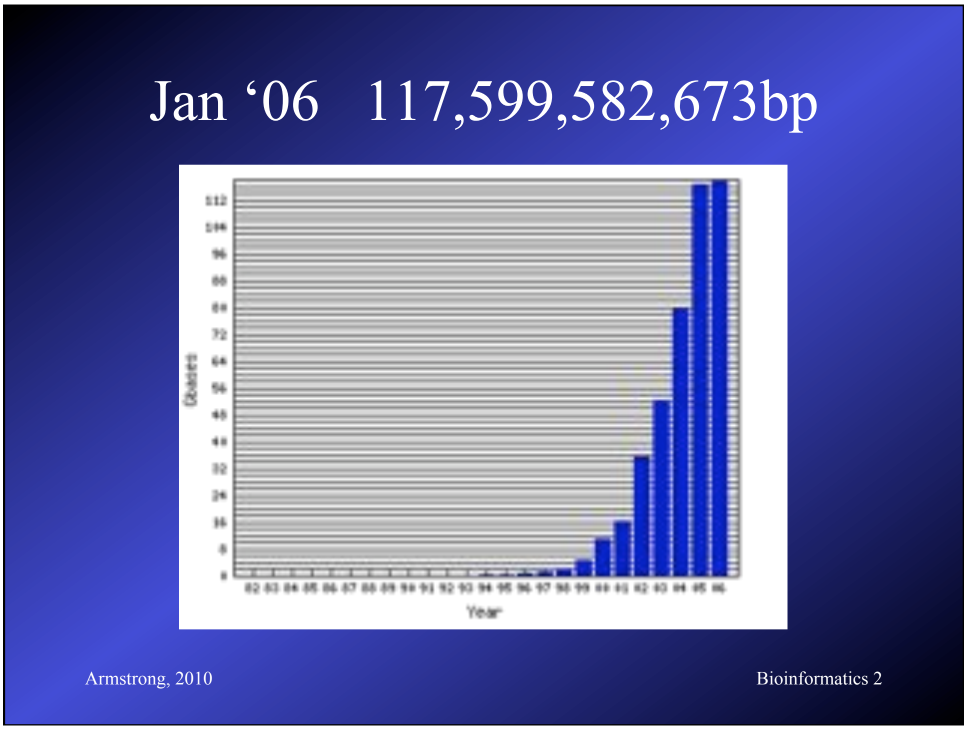
Limited annotation

- Unique accession number
- Submitting author(s)
- Brief annotation if available
- Source (cDNA, EST, genomic etc)
- Species
- Reference or Patent details



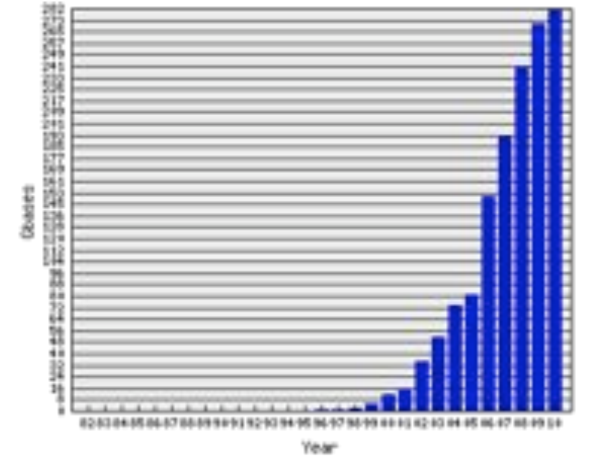
EMBL file tags

```
ID - identification              (begins each entry; 1 per entry)
AC - accession number            (>=1 per entry)
SV - new sequence identifier     (>=1 per entry)
DT - date                        (2 per entry)
DE - description                 (>=1 per entry)
KW - keyword                     (>=1 per entry)
OS - organism species            (>=1 per entry)
OC - organism classification     (>=1 per entry)
OG - organelle                   (0 or 1 per entry)
RN - reference number            (>=1 per entry)
RC - reference comment           (>=0 per entry)
RP - reference positions         (>=1 per entry)
RX - reference cross-reference   (>=0 per entry)
RA - reference author(s)         (>=1 per entry)
RT - reference title             (>=1 per entry)
RL - reference location          (>=1 per entry)
DR - database cross-reference    (>=0 per entry)
FH - feature table header        (0 or 2 per entry)
FT - feature table data          (>=0 per entry)
CC - comments or notes           (>=0 per entry)
XX - spacer line                 (many per entry)
SQ - sequence header             (1 per entry)
bb - (blanks) sequence data      (>=1 per entry)
// - termination line            (ends each entry; 1 per entry)
```

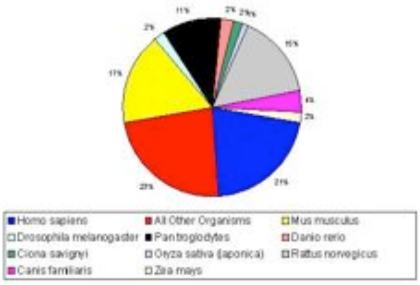


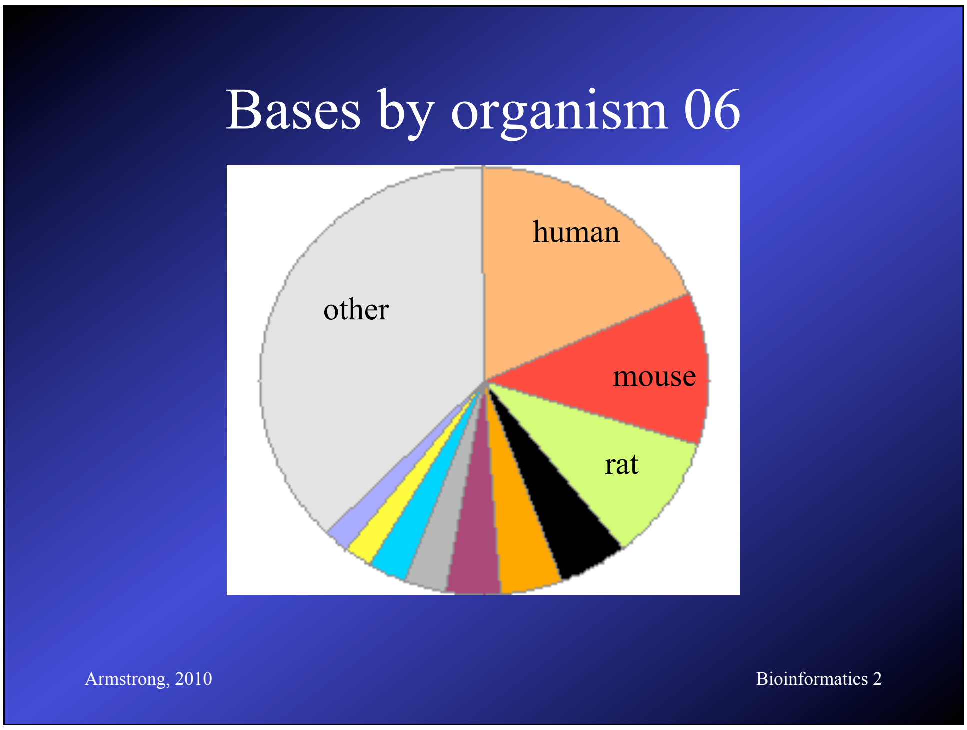
Jan '06 117,599,582,673bp

Feb'10 281,244,445,986bp



Bases by organism 04

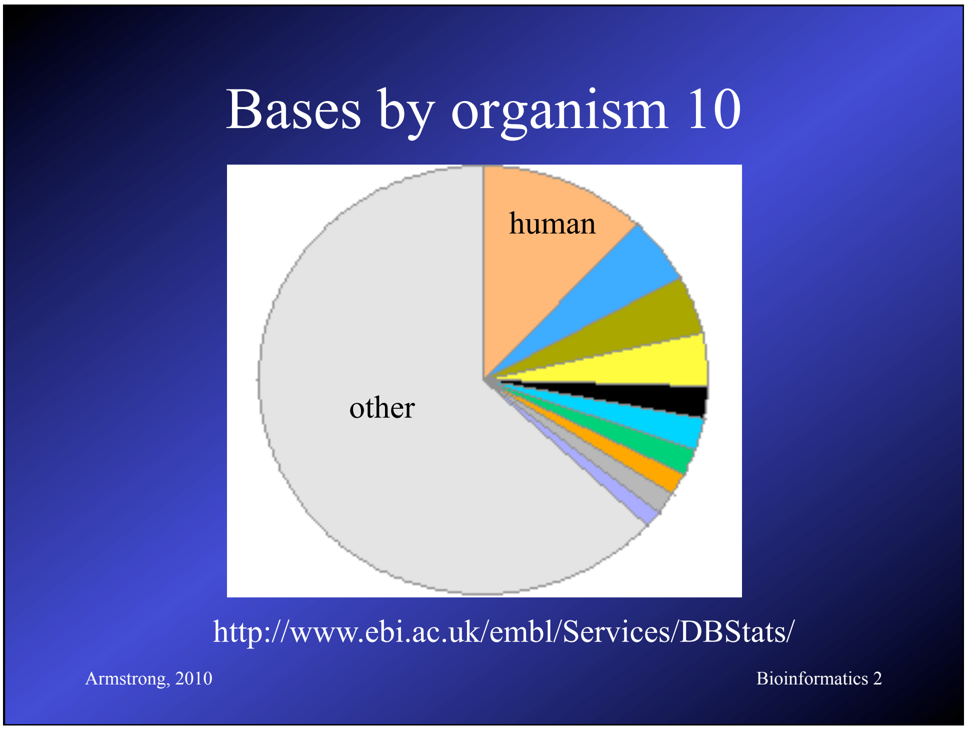




Bases by organism 06

Armstrong, 2010

Bioinformatics 2

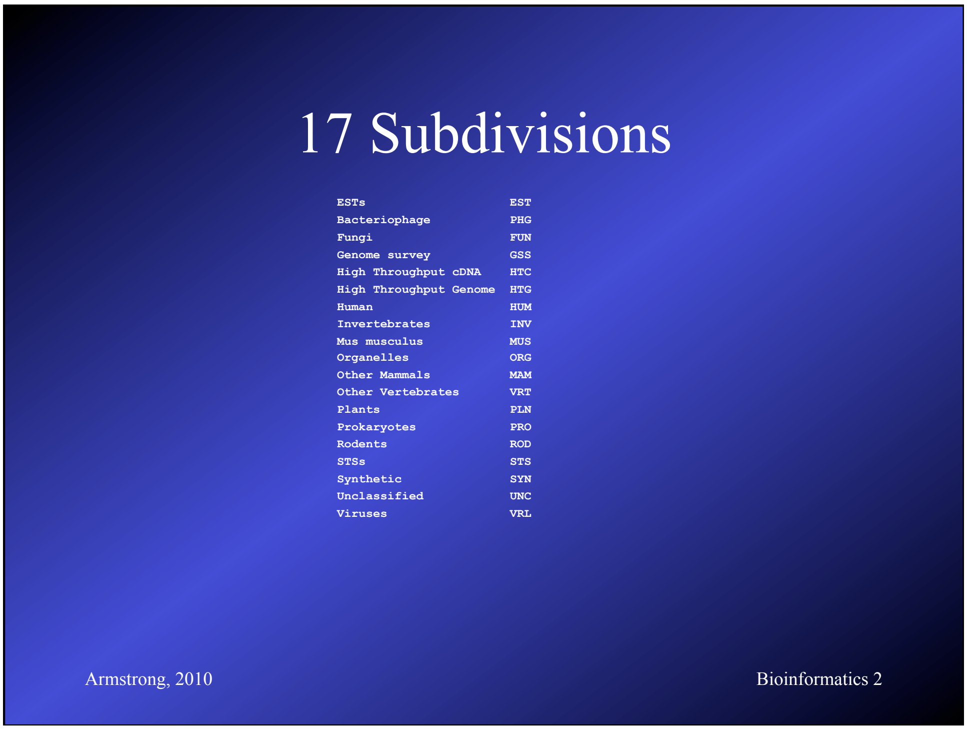


Bases by organism 10

http://www.ebi.ac.uk/embl/Services/DBStats/

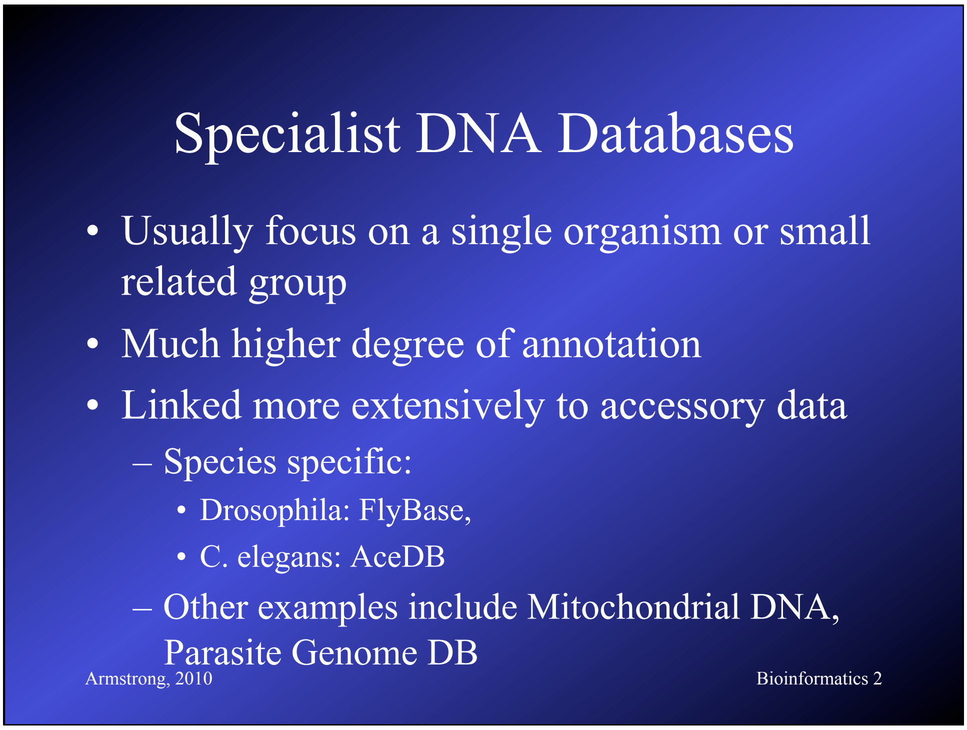
Armstrong, 2010

Bioinformatics 2



17 Subdivisions

ESTs	EST
Bacteriophage	PHG
Fungi	FUN
Genome survey	GSS
High Throughput cDNA	HTC
High Throughput Genome	HTG
Human	HUM
Invertebrates	INV
Mus musculus	MUS
Organelles	ORG
Other Mammals	MAM
Other Vertebrates	VRT
Plants	PLN
Prokaryotes	PRO
Rodents	ROD
STSs	STS
Synthetic	SYN
Unclassified	UNC
Viruses	VRL



Specialist DNA Databases

- Usually focus on a single organism or small related group
- Much higher degree of annotation
- Linked more extensively to accessory data
  - Species specific:
    - Drosophila: FlyBase,
    - C. elegans: AceDB
  - Other examples include Mitochondrial DNA, Parasite Genome DB

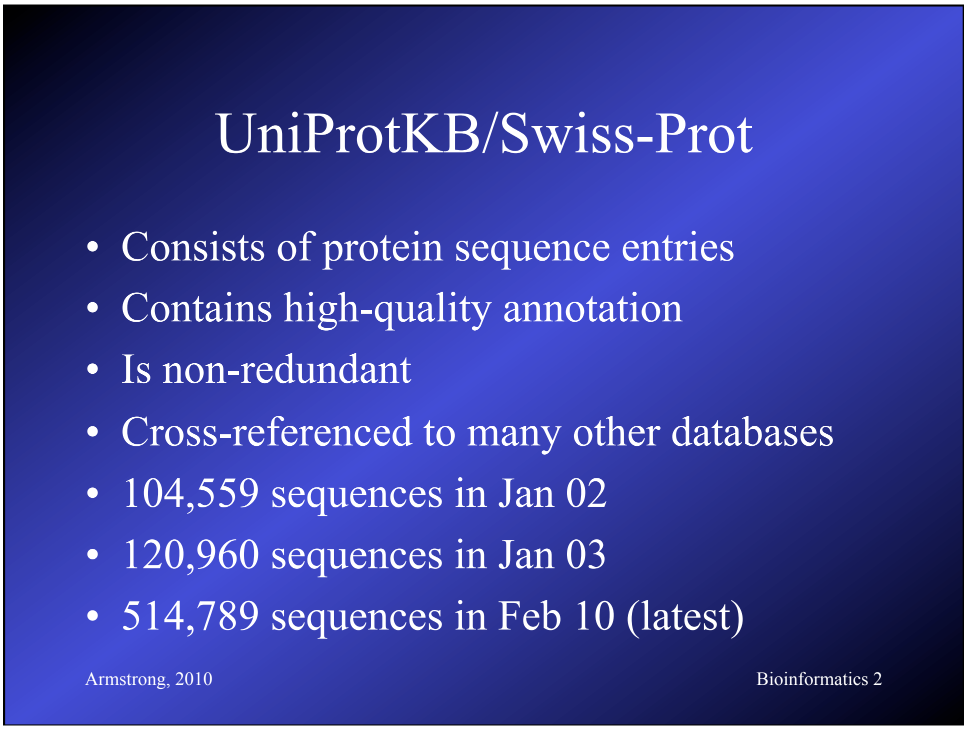
FlyBase

flybase.bio.indiana.edu

- Includes the entire annotated genome searchable by BLAST or by text queries
- Also includes a detailed ontology or standard nomenclature for *Drosophila*
- Also provides information on all literature, researchers, mutations, genetic stocks and technical resources.
- Full mirror at EBI

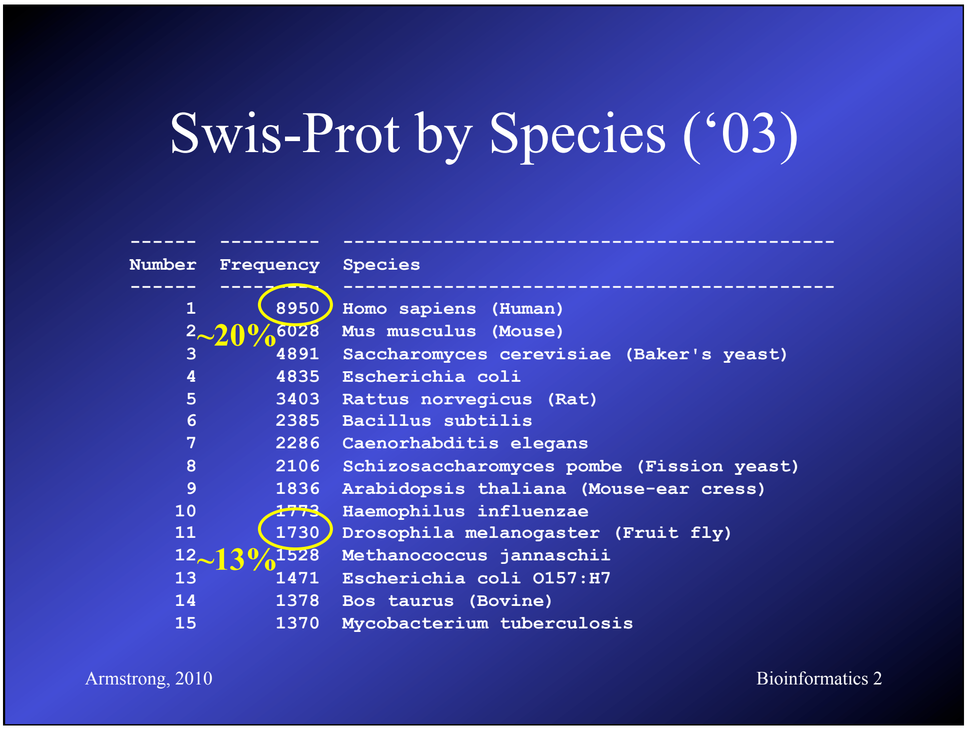
Protein DBs

- Primary Sequence DBs
  - UniProt, TrEMBL, GenPept
- Protein Structure DBs
  - PDB, MSD
- Protein Domain Homology DBs
  - InterPro, CluSTr



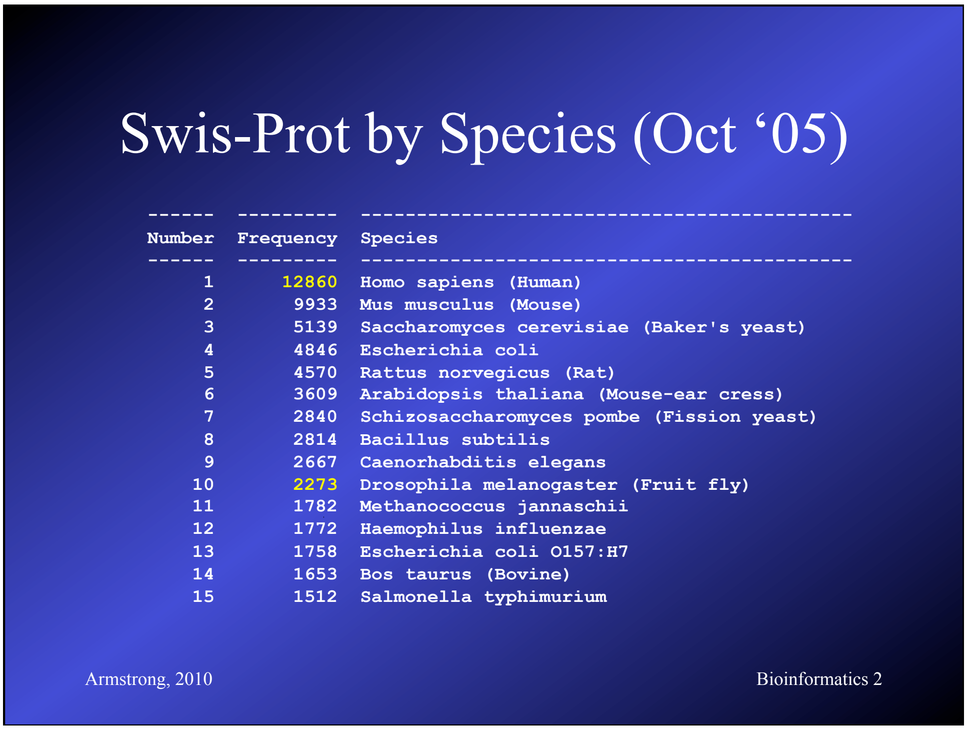
UniProtKB/Swiss-Prot

- Consists of protein sequence entries
- Contains high-quality annotation
- Is non-redundant
- Cross-referenced to many other databases
- 104,559 sequences in Jan 02
- 120,960 sequences in Jan 03
- 514,789 sequences in Feb 10 (latest)



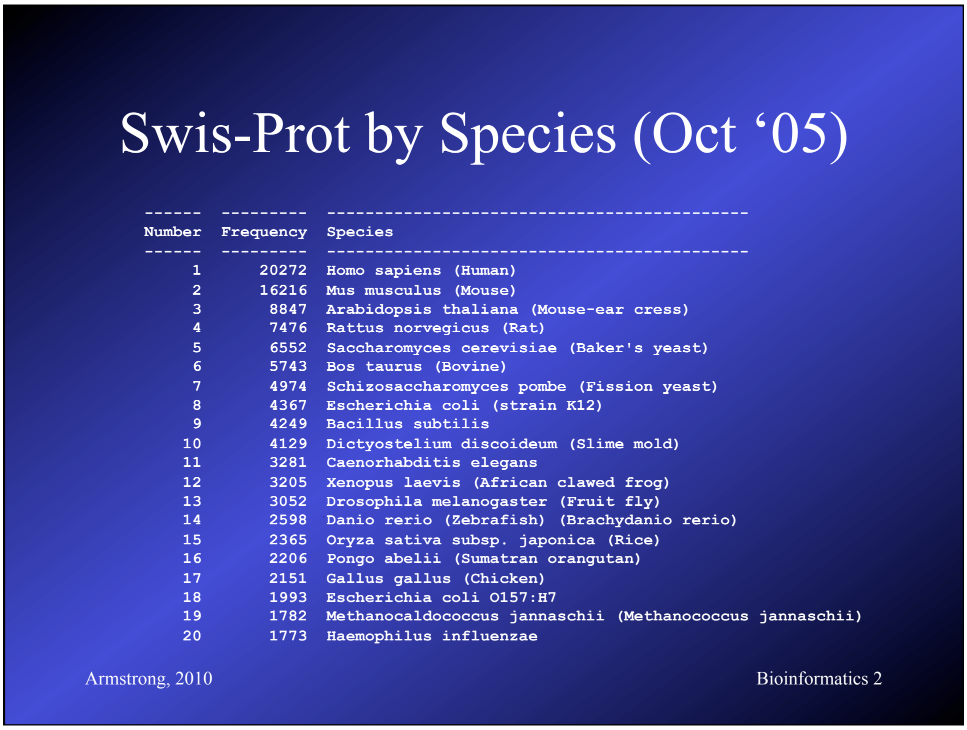
Swis-Prot by Species (‘03)

```
------  ---------  --------------------------------------------
Number  Frequency  Species
------  ---------  --------------------------------------------
     1       8950  Homo sapiens (Human)
     2~20%   6028  Mus musculus (Mouse)
     3       4891  Saccharomyces cerevisiae (Baker's yeast)
     4       4835  Escherichia coli
     5       3403  Rattus norvegicus (Rat)
     6       2385  Bacillus subtilis
     7       2286  Caenorhabditis elegans
     8       2106  Schizosaccharomyces pombe (Fission yeast)
     9       1836  Arabidopsis thaliana (Mouse-ear cress)
    10       1773  Haemophilus influenzae
    11       1730  Drosophila melanogaster (Fruit fly)
    12~13%   1528  Methanococcus jannaschii
    13       1471  Escherichia coli O157:H7
    14       1378  Bos taurus (Bovine)
    15       1370  Mycobacterium tuberculosis
```



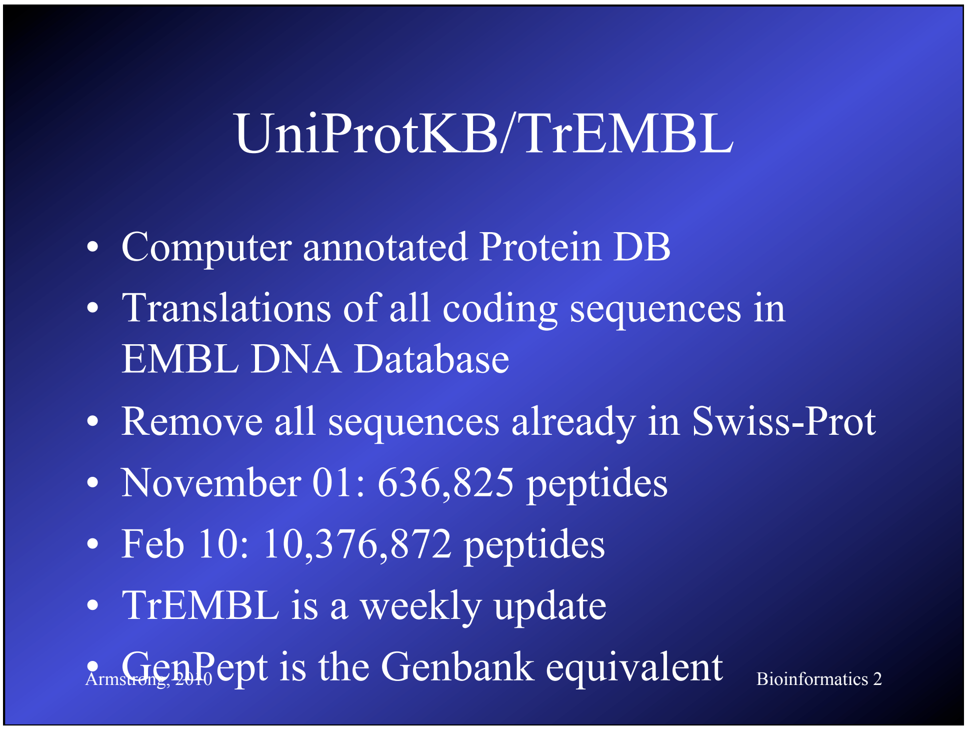
Swis-Prot by Species (Oct ‘05)

```
------  ---------  ---------------------------------------------
Number  Frequency  Species
------  ---------  ---------------------------------------------
     1      12860  Homo sapiens (Human)
     2       9933  Mus musculus (Mouse)
     3       5139  Saccharomyces cerevisiae (Baker's yeast)
     4       4846  Escherichia coli
     5       4570  Rattus norvegicus (Rat)
     6       3609  Arabidopsis thaliana (Mouse-ear cress)
     7       2840  Schizosaccharomyces pombe (Fission yeast)
     8       2814  Bacillus subtilis
     9       2667  Caenorhabditis elegans
    10       2273  Drosophila melanogaster (Fruit fly)
    11       1782  Methanococcus jannaschii
    12       1772  Haemophilus influenzae
    13       1758  Escherichia coli O157:H7
    14       1653  Bos taurus (Bovine)
    15       1512  Salmonella typhimurium
```



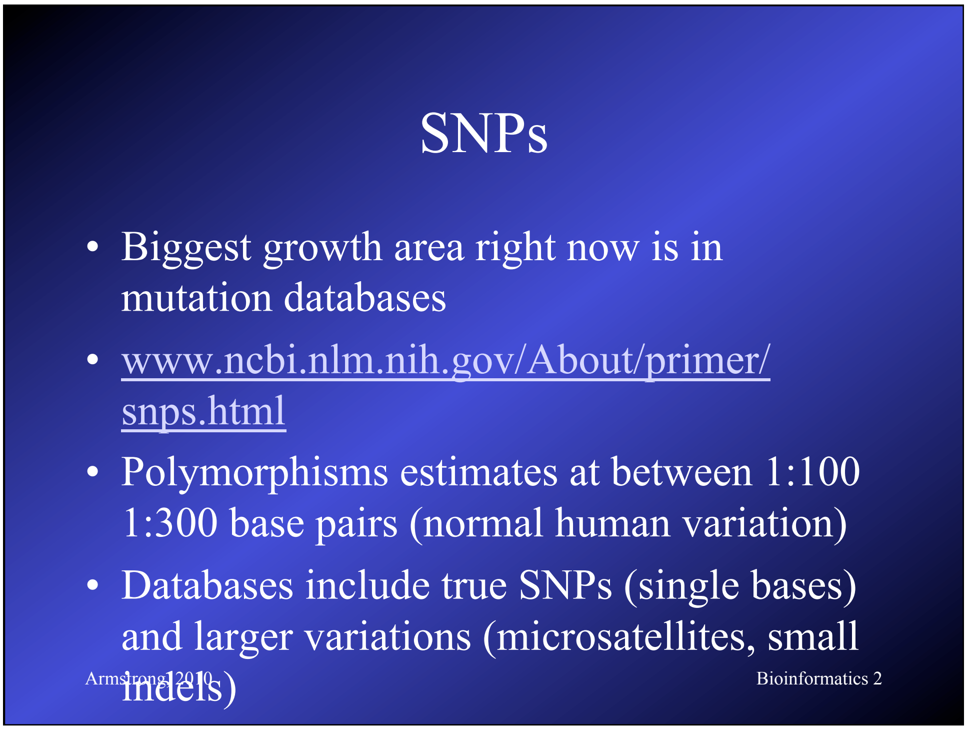
Swis-Prot by Species (Oct ‘05)

```
------  ---------  ---------------------------------------------
Number  Frequency  Species
------  ---------  ---------------------------------------------
     1      20272  Homo sapiens (Human)
     2      16216  Mus musculus (Mouse)
     3       8847  Arabidopsis thaliana (Mouse-ear cress)
     4       7476  Rattus norvegicus (Rat)
     5       6552  Saccharomyces cerevisiae (Baker's yeast)
     6       5743  Bos taurus (Bovine)
     7       4974  Schizosaccharomyces pombe (Fission yeast)
     8       4367  Escherichia coli (strain K12)
     9       4249  Bacillus subtilis
    10       4129  Dictyostelium discoideum (Slime mold)
    11       3281  Caenorhabditis elegans
    12       3205  Xenopus laevis (African clawed frog)
    13       3052  Drosophila melanogaster (Fruit fly)
    14       2598  Danio rerio (Zebrafish) (Brachydanio rerio)
    15       2365  Oryza sativa subsp. japonica (Rice)
    16       2206  Pongo abelii (Sumatran orangutan)
    17       2151  Gallus gallus (Chicken)
    18       1993  Escherichia coli O157:H7
    19       1782  Methanocaldococcus jannaschii (Methanococcus jannaschii)
    20       1773  Haemophilus influenzae
```



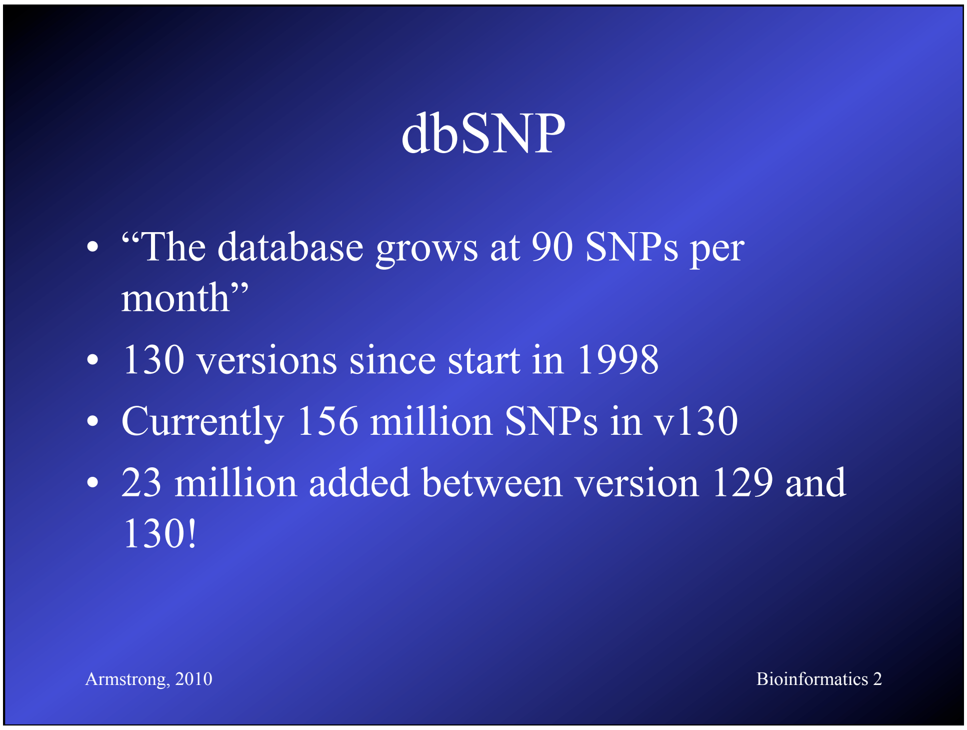
UniProtKB/TrEMBL

- Computer annotated Protein DB
- Translations of all coding sequences in EMBL DNA Database
- Remove all sequences already in Swiss-Prot
- November 01: 636,825 peptides
- Feb 10: 10,376,872 peptides
- TrEMBL is a weekly update
- GenPept is the Genbank equivalent



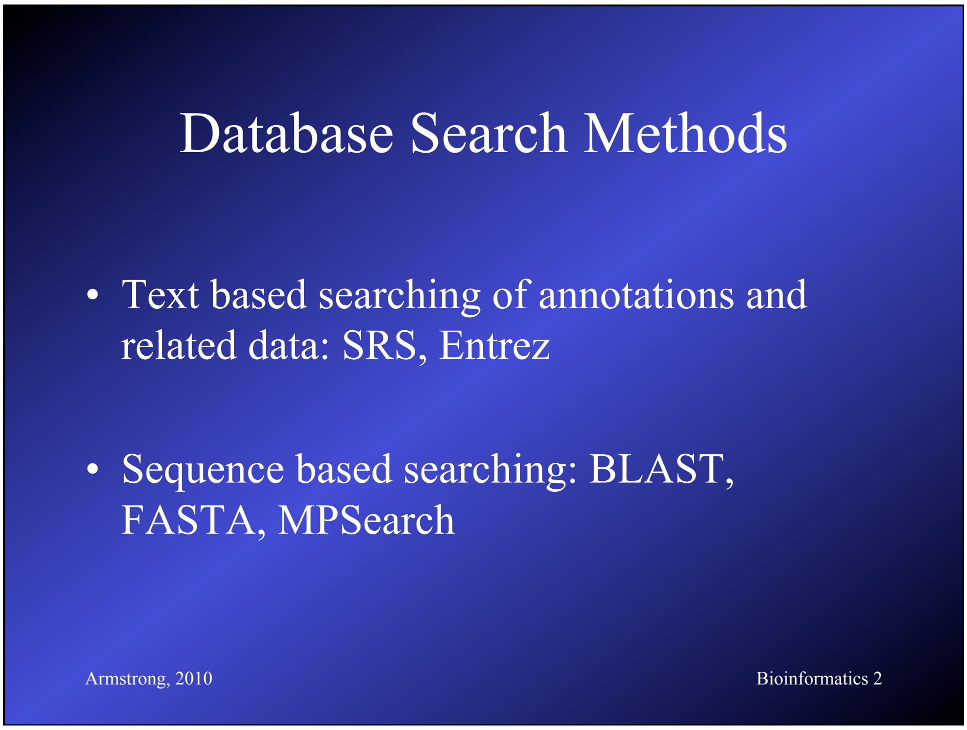
SNPs

- Biggest growth area right now is in mutation databases
- www.ncbi.nlm.nih.gov/About/primer/snps.html
- Polymorphisms estimates at between 1:100 1:300 base pairs (normal human variation)
- Databases include true SNPs (single bases) and larger variations (microsatellites, small indels)



dbSNP

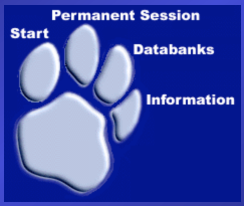
- “The database grows at 90 SNPs per month”
- 130 versions since start in 1998
- Currently 156 million SNPs in v130
- 23 million added between version 129 and 130!



Database Search Methods

- Text based searching of annotations and related data: SRS, Entrez
- Sequence based searching: BLAST, FASTA, MPSearch

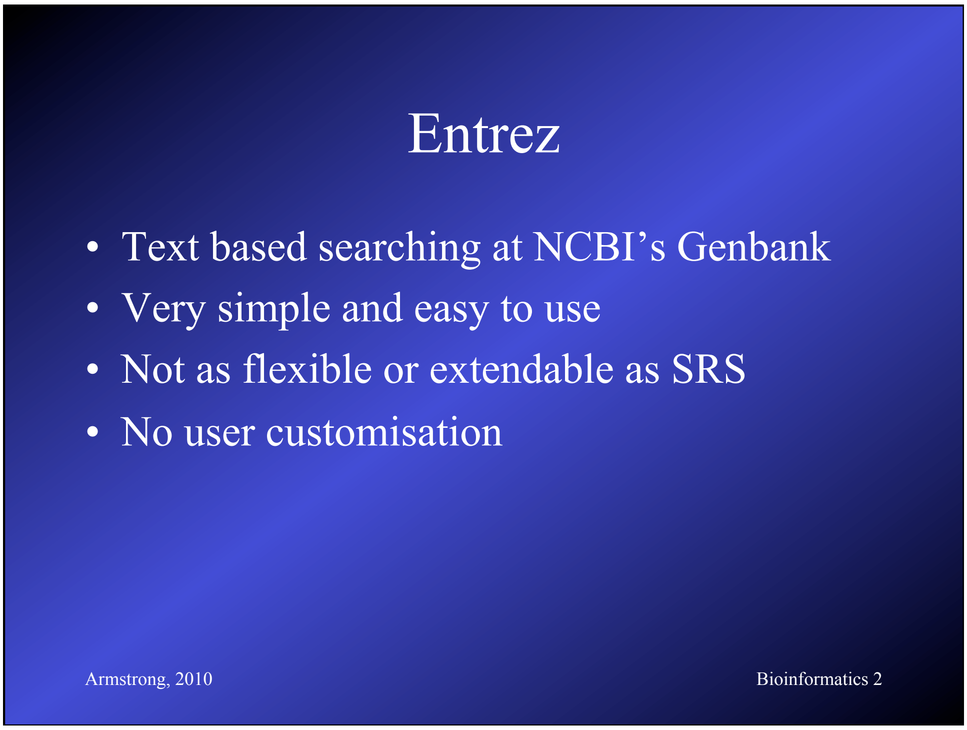
SRS



- Sequence Retrieval System
  - Powerful search of EMBL annotation
  - Linked to over 80 other data sources
  - Also includes results from automated searches

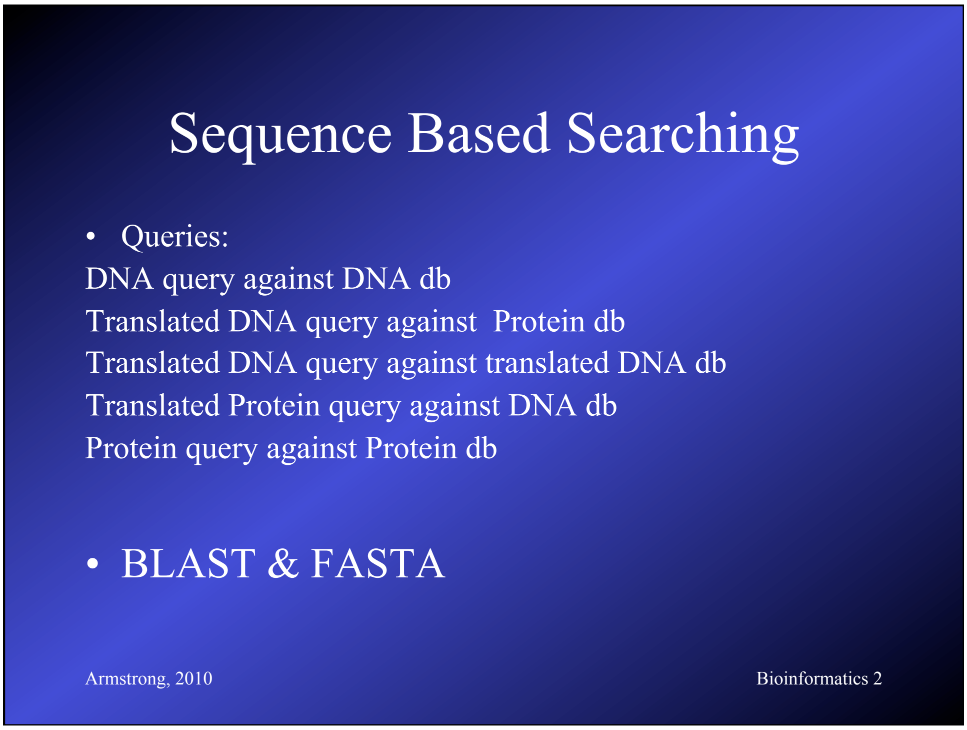
SRS data sources

- Primary Sequence: EMBL, SwissProt
- References/Literature: Medline
- Protein Homology: Prosite, Prints
- Sequence Related: Blocks, UTR, Taxonomy
- Transcription Factor: TFACTOR, TFSITE
- Search Results: BLAST, FASTA, CLUSTALW
- Protein Structure: PDB
- Also, Mutations, Pathways, other specialist DBs



Entrez

- Text based searching at NCBI’s Genbank
- Very simple and easy to use
- Not as flexible or extendable as SRS
- No user customisation

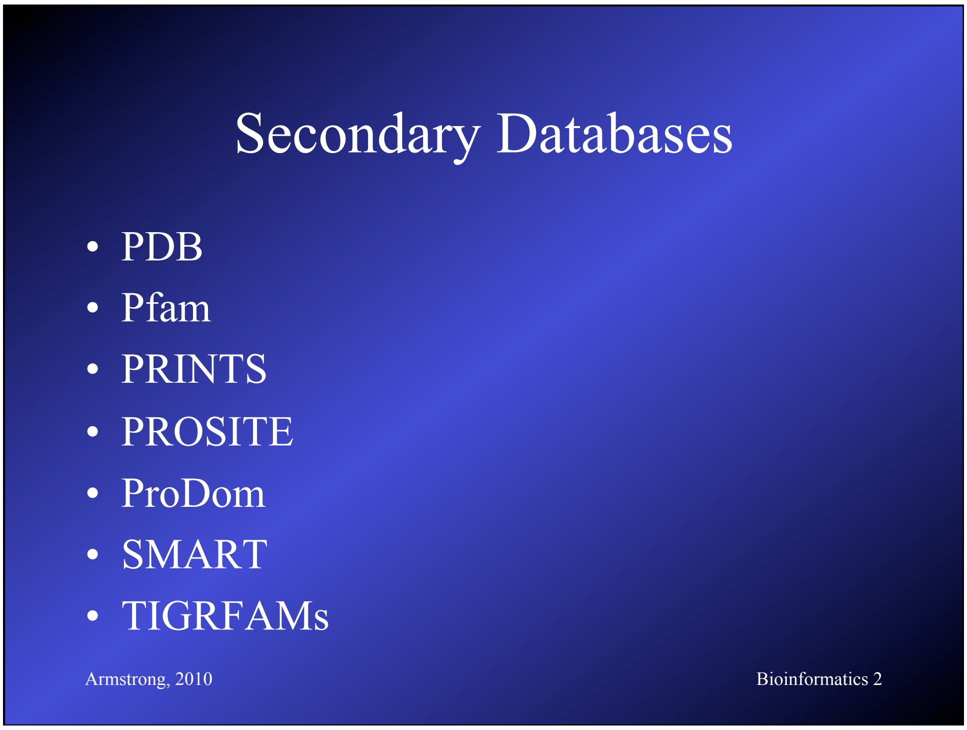


Sequence Based Searching

- Queries:

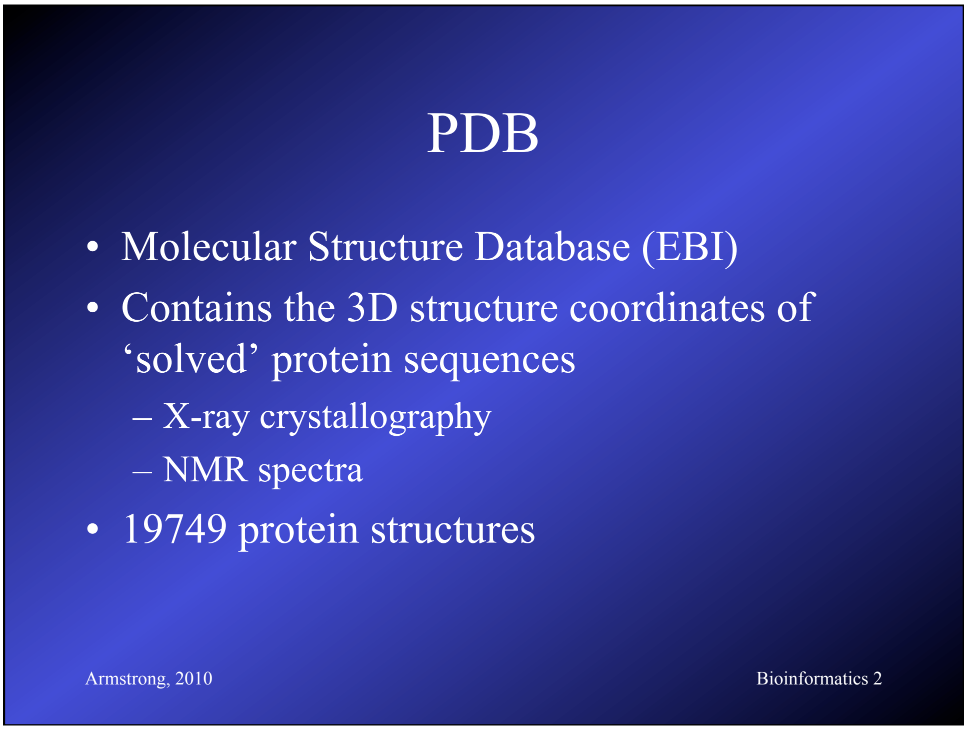
DNA query against DNA db
Translated DNA query against Protein db
Translated DNA query against translated DNA db
Translated Protein query against DNA db
Protein query against Protein db

- BLAST & FASTA



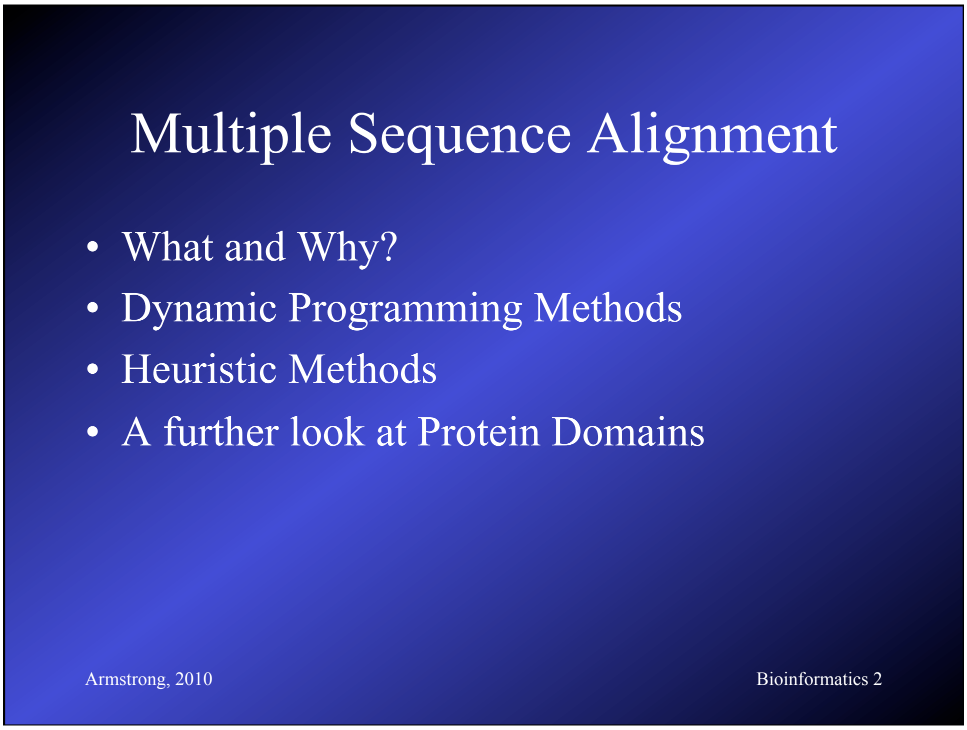
Secondary Databases

- PDB
- Pfam
- PRINTS
- PROSITE
- ProDom
- SMART
- TIGRFAMs



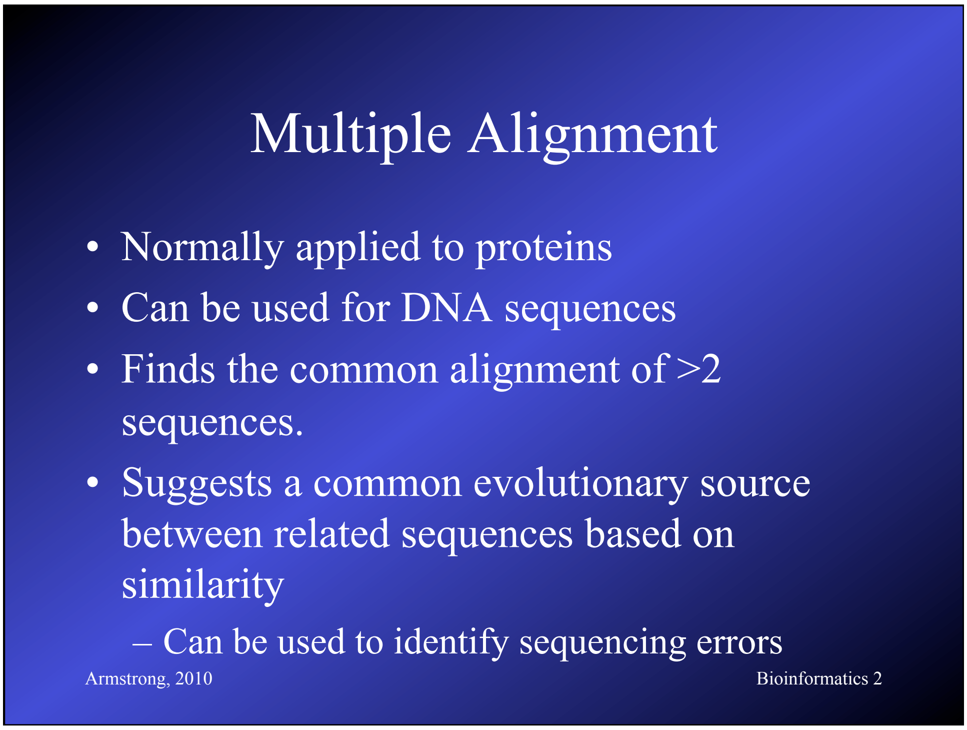
PDB

- Molecular Structure Database (EBI)
- Contains the 3D structure coordinates of ‘solved’ protein sequences
  - X-ray crystallography
  - NMR spectra
- 19749 protein structures



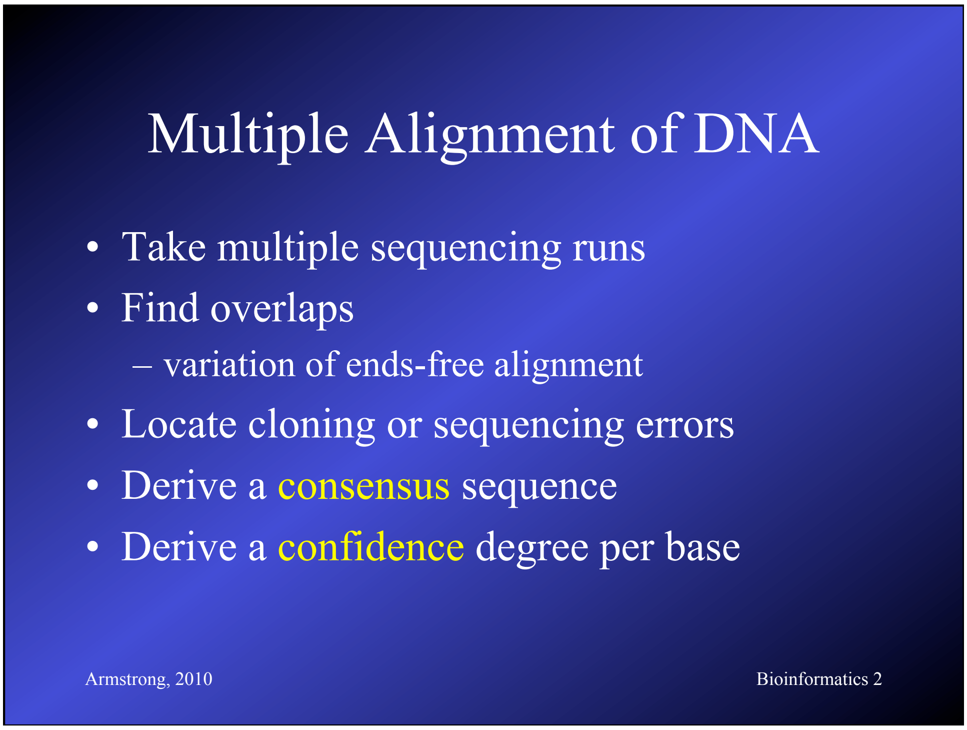
Multiple Sequence Alignment

- What and Why?
- Dynamic Programming Methods
- Heuristic Methods
- A further look at Protein Domains



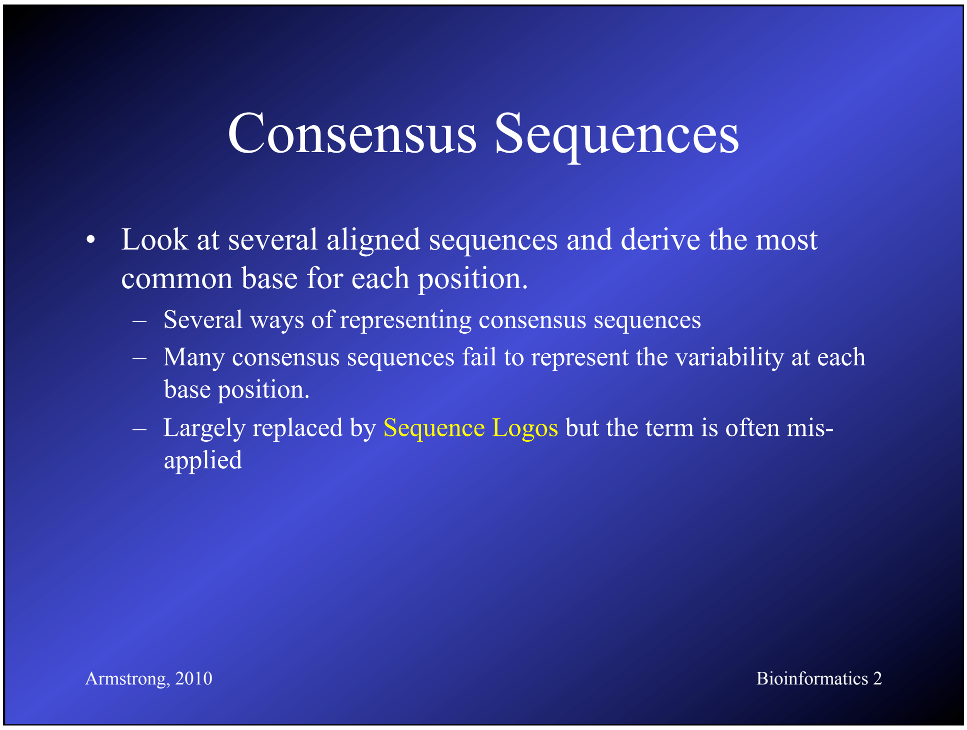
Multiple Alignment

- Normally applied to proteins
- Can be used for DNA sequences
- Finds the common alignment of >2 sequences.
- Suggests a common evolutionary source between related sequences based on similarity
  - Can be used to identify sequencing errors



Multiple Alignment of DNA

- Take multiple sequencing runs
- Find overlaps
  - variation of ends-free alignment
- Locate cloning or sequencing errors
- Derive a consensus sequence
- Derive a confidence degree per base



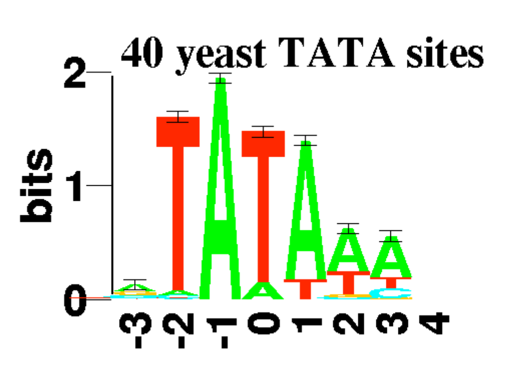
Consensus Sequences

- Look at several aligned sequences and derive the most common base for each position.
  - Several ways of representing consensus sequences
  - Many consensus sequences fail to represent the variability at each base position.
  - Largely replaced by Sequence Logos but the term is often mis-applied

Sequence Logos

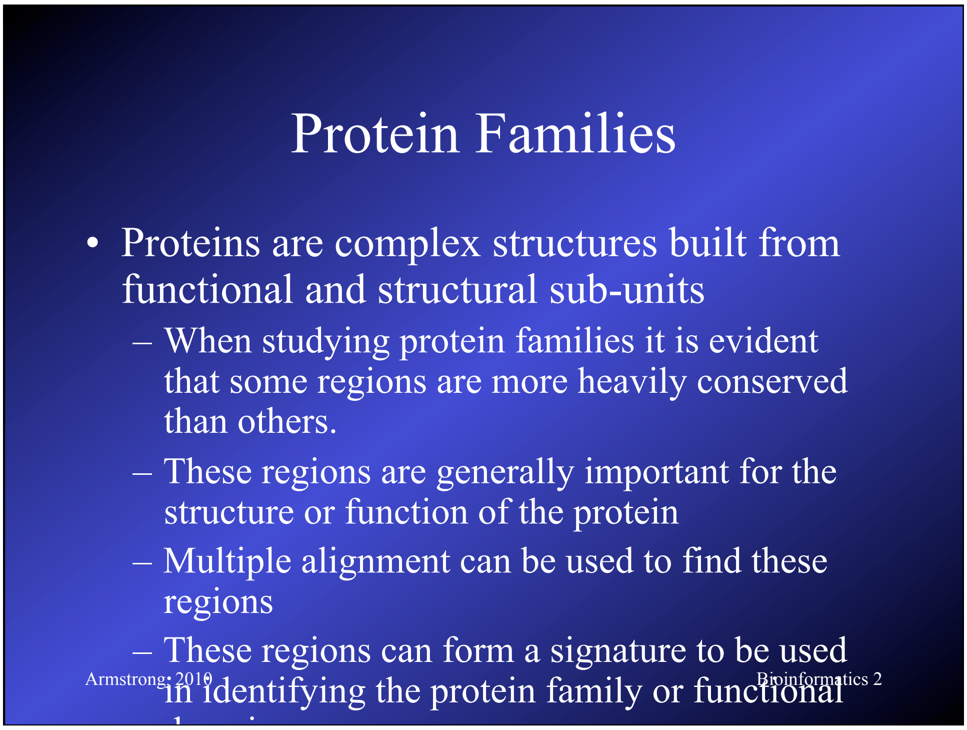
- Example, from an alignment of the TATA box in yeast genes:

We now have a confidence level for each base at each position



Multiple Alignment of Proteins

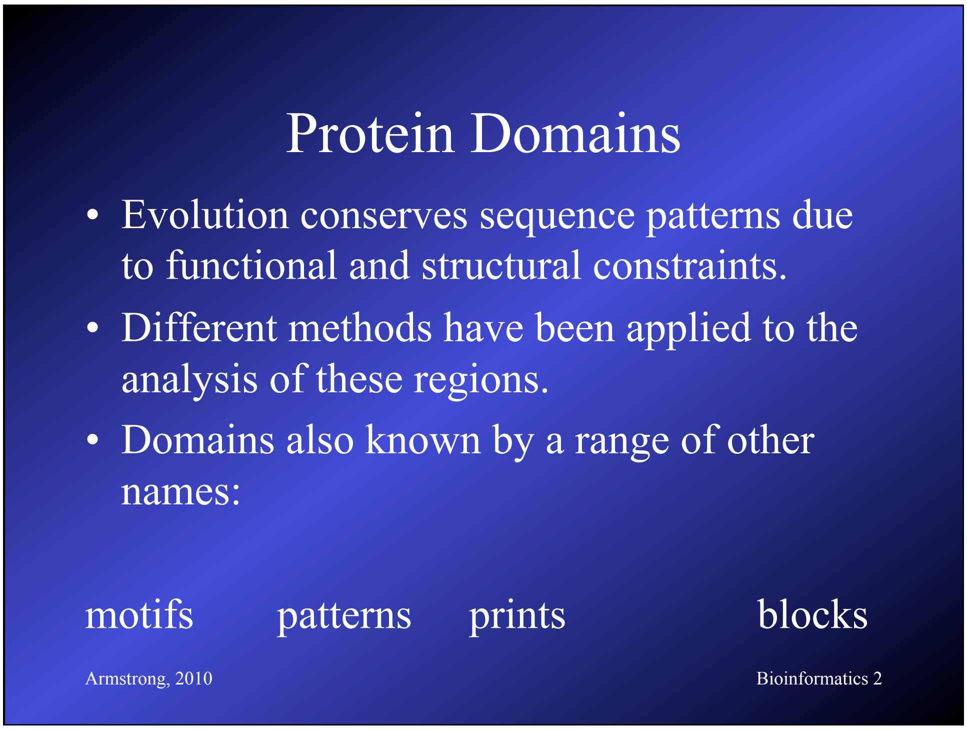
- Multiple Alignment of Proteins
- Identify Protein Families
- Find conserved Protein Domains
- Predict evolutionary precursor sequences
- Predict evolutionary trees



Protein Families

- Proteins are complex structures built from functional and structural sub-units
  - When studying protein families it is evident that some regions are more heavily conserved than others.
  - These regions are generally important for the structure or function of the protein
  - Multiple alignment can be used to find these regions
  - These regions can form a signature to be used in identifying the protein family or functional

Armstrong, 2010 Bioinformatics 2

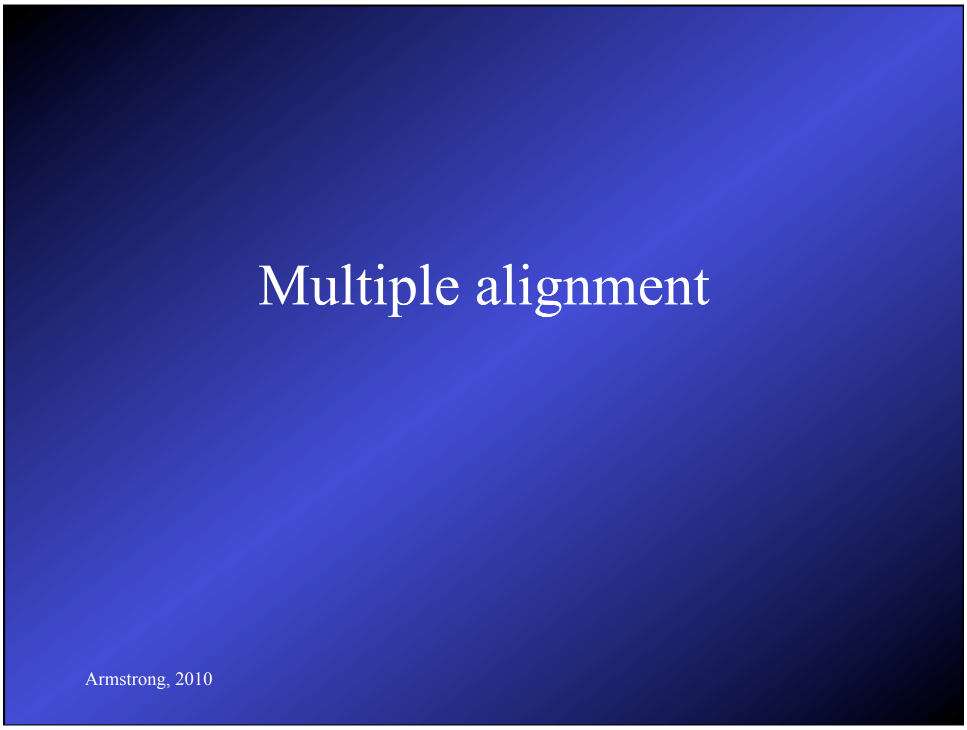


Protein Domains

- Evolution conserves sequence patterns due to functional and structural constraints.
- Different methods have been applied to the analysis of these regions.
- Domains also known by a range of other names:

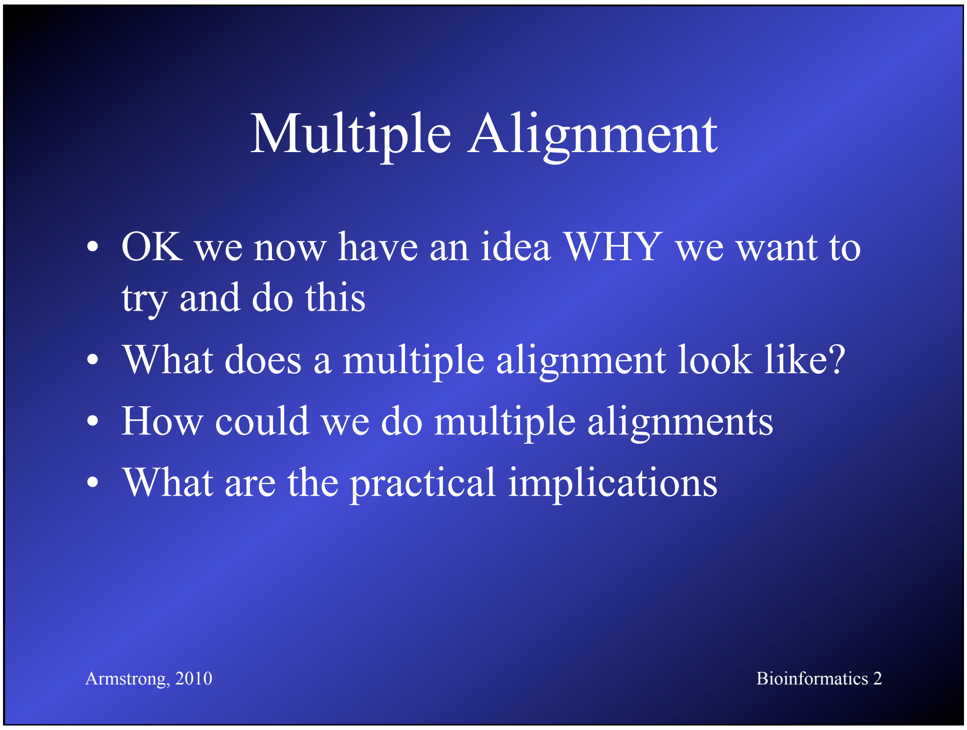
motifs patterns prints blocks

Armstrong, 2010 Bioinformatics 2



Multiple alignment

Armstrong, 2010



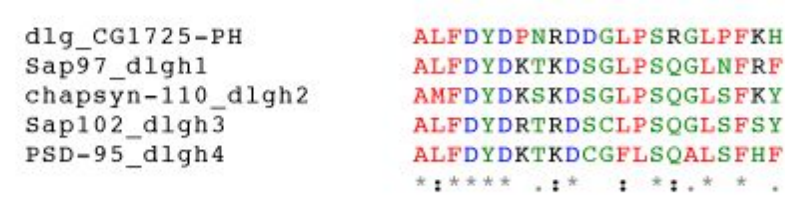
Multiple Alignment

- OK we now have an idea WHY we want to try and do this
- What does a multiple alignment look like?
- How could we do multiple alignments
- What are the practical implications

Armstrong, 2010

Bioinformatics 2

Multiple alignment table



```
dlg_CG1725-PH            ALFDYDPNRDDGLPSRGLPFKH
Sap97_dlgh1              ALFDYDKTKDSGLPSQGLNFRF
chapsyn-110_dlgh2        AMFDYDKSKDSGLPSQGLSFKY
Sap102_dlgh3             ALFDYDRTRDSCLPSQGLSFSY
PSD-95_dlgh4             ALFDYDKTKDCGFLSQALSFHF
                         *:**** .:*  : *:.* * .
```

A consensus character is the one that minimises the distance between it and all the other characters in the column

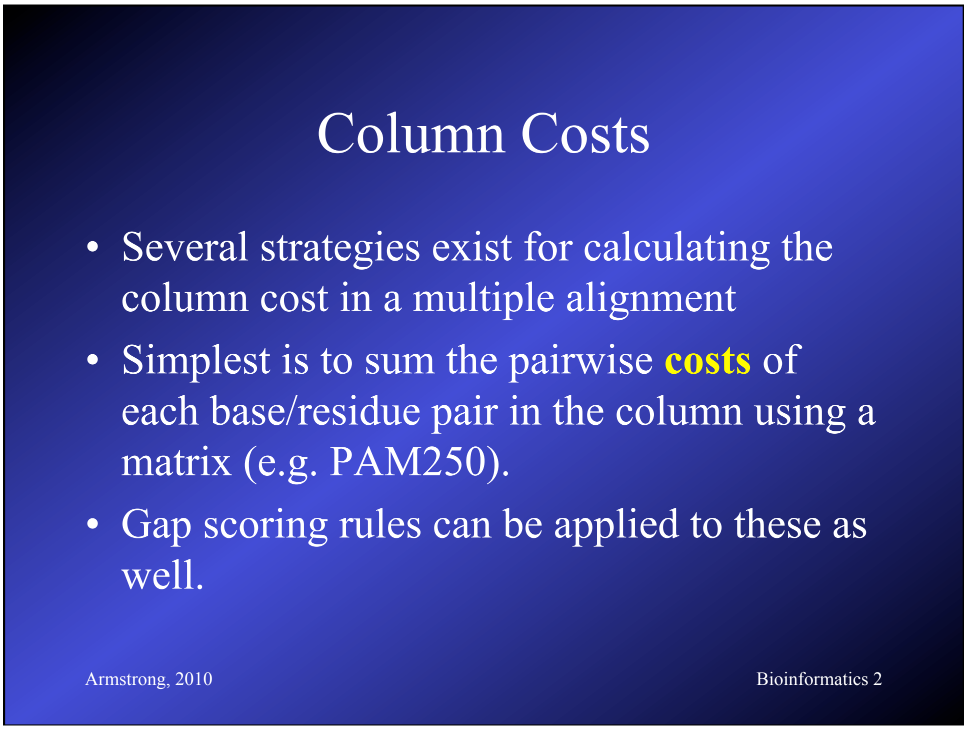
Conservatived or Identical residues are colour coded

Scoring Multiple Alignments

- We need to score on columns with more than 2 bases or residues:

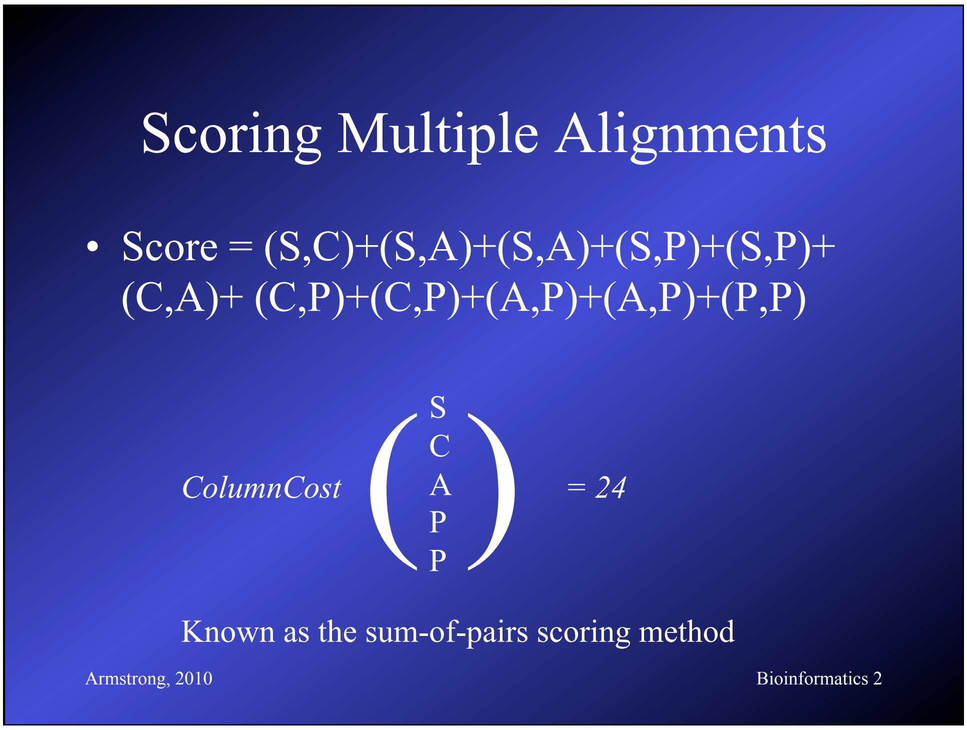
$$\mathit{ColumnCost}\begin{pmatrix}S\\C\\A\\P\\P\end{pmatrix} = 24$$

Multiple alignments are usually scored on cost/difference rather than similarity



Column Costs

- Several strategies exist for calculating the column cost in a multiple alignment
- Simplest is to sum the pairwise **costs** of each base/residue pair in the column using a matrix (e.g. PAM250).
- Gap scoring rules can be applied to these as well.

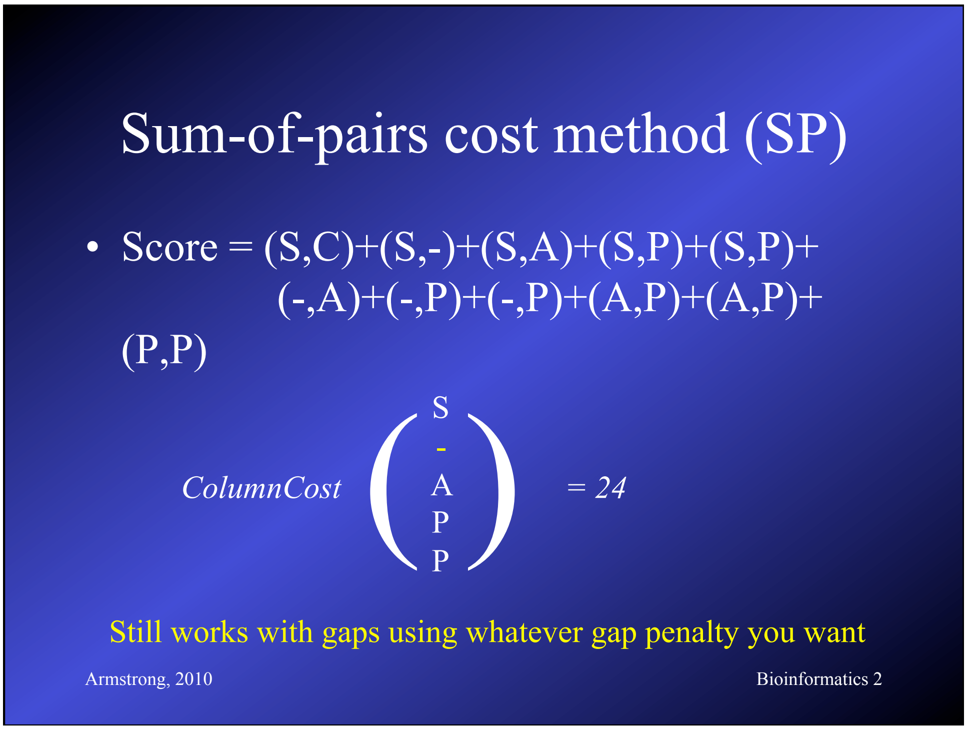


Scoring Multiple Alignments

- Score = (S,C)+(S,A)+(S,A)+(S,P)+(S,P)+(C,A)+ (C,P)+(C,P)+(A,P)+(A,P)+(P,P)

$$\mathit{ColumnCost}\begin{pmatrix}S\\C\\A\\P\\P\end{pmatrix} = 24$$

Known as the sum-of-pairs scoring method



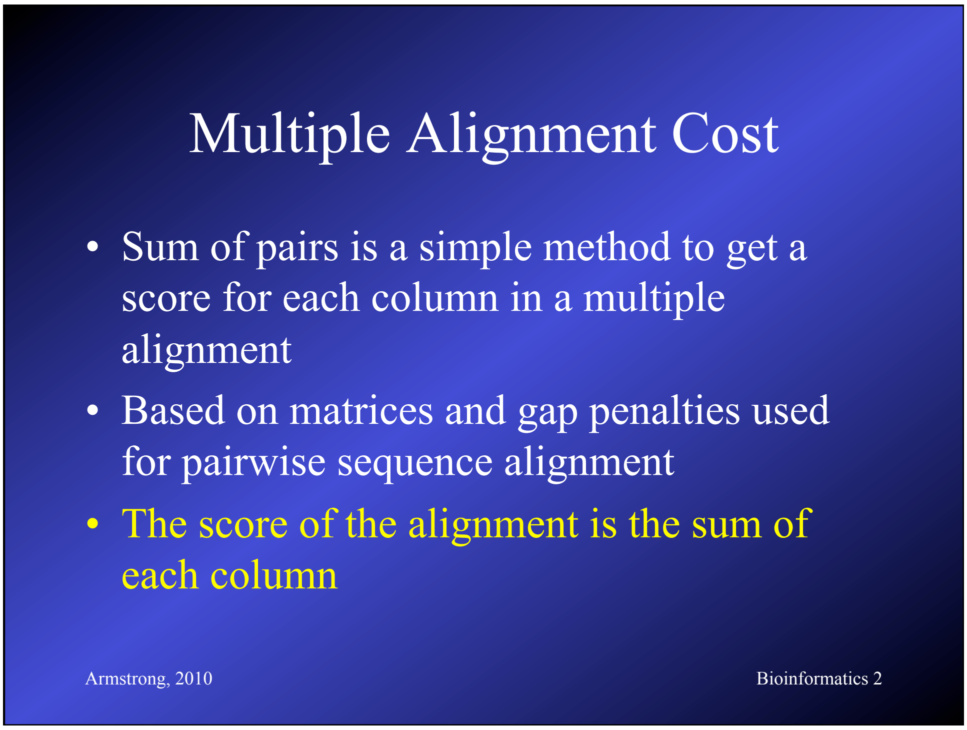
Sum-of-pairs cost method (SP)

- Score = (S,C)+(S,-)+(S,A)+(S,P)+(S,P)+ (-,A)+(-,P)+(-,P)+(A,P)+(A,P)+ (P,P)

$$\mathit{ColumnCost}\begin{pmatrix}\mathrm{S}\\ -\\ \mathrm{A}\\ \mathrm{P}\\ \mathrm{P}\end{pmatrix} = 24$$

Still works with gaps using whatever gap penalty you want

Armstrong, 2010 Bioinformatics 2



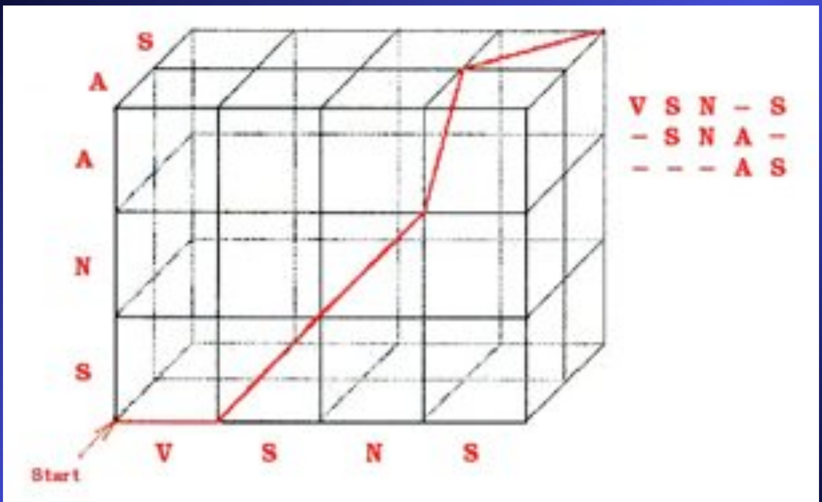
Multiple Alignment Cost

- Sum of pairs is a simple method to get a score for each column in a multiple alignment
- Based on matrices and gap penalties used for pairwise sequence alignment
- The score of the alignment is the sum of each column

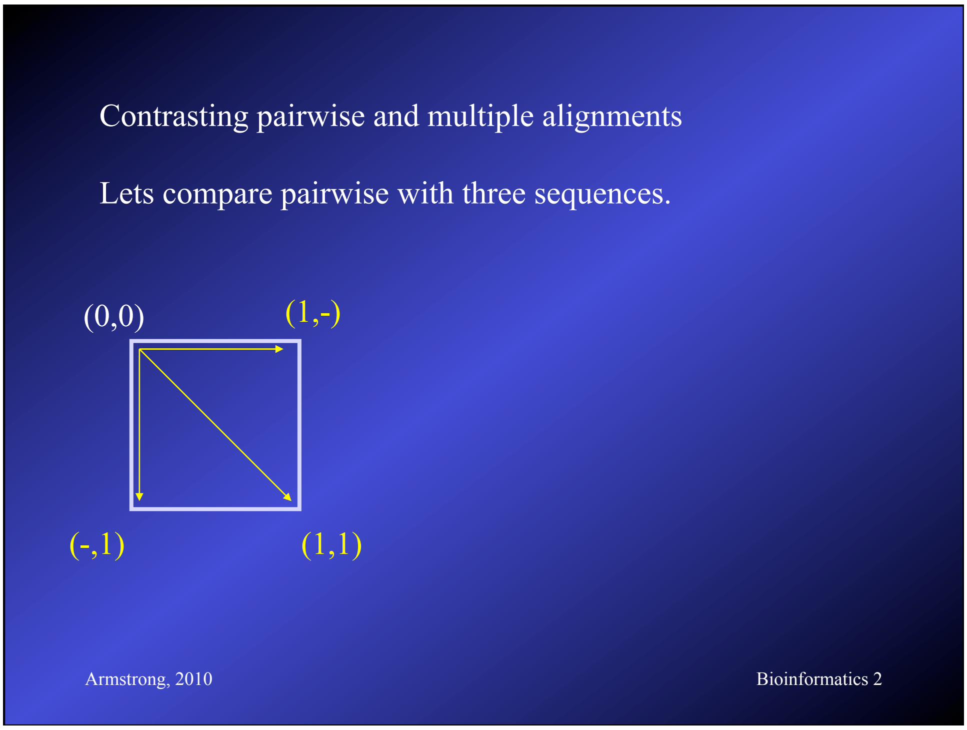
Armstrong, 2010 Bioinformatics 2

Optimal Multiple Alignment

- The best alignment is generally the one with the lowest score (i.e. least difference)
  - depends on the scoring rules used.
- Like pairwise cases, each alignment represents a path through a matrix
- For multiple alignment, the matrix is *n*-dimensional
  - where *n*=number of sequences



(Murata, Richardson and Sussman 1999)

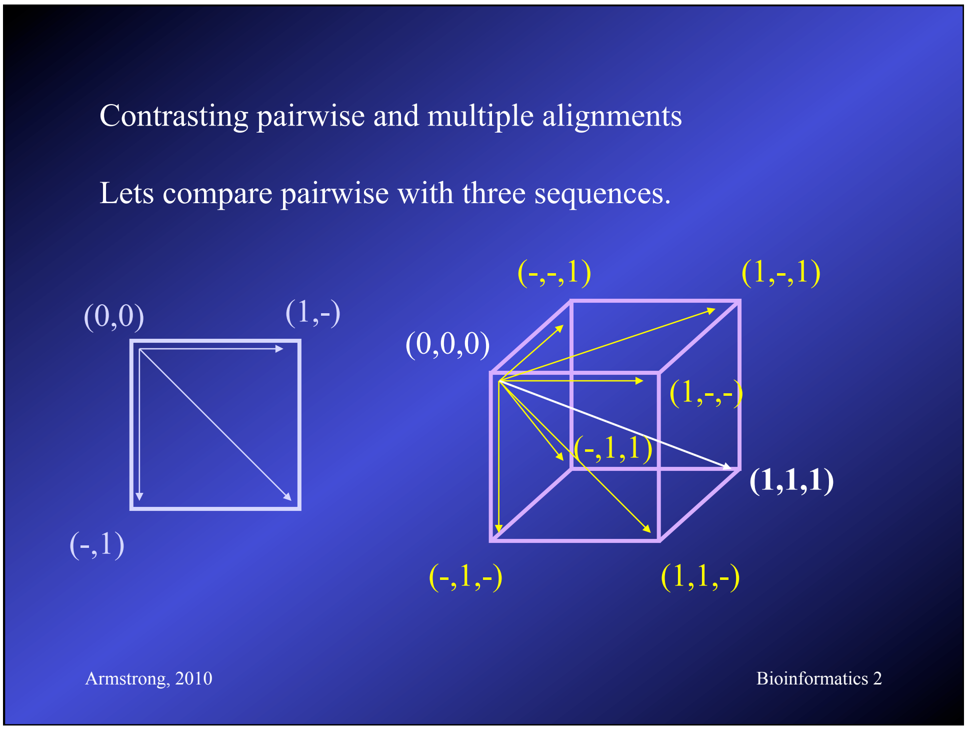


Contrasting pairwise and multiple alignments

Lets compare pairwise with three sequences.

Armstrong, 2010

Bioinformatics 2



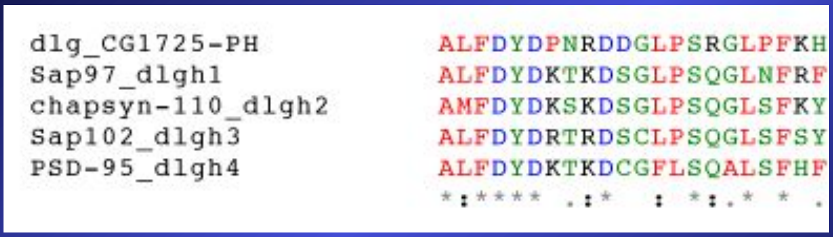
Contrasting pairwise and multiple alignments

Lets compare pairwise with three sequences.

Armstrong, 2010

Bioinformatics 2

Multiple alignment table



```
dlg_CG1725-PH           ALFDYDPNRDDGLPSRGLPFKH
Sap97_dlgh1             ALFDYDKTKDSGLPSQGLNFRF
chapsyn-110_dlgh2       AMFDYDKSKDSGLPSQGLSFKY
Sap102_dlgh3            ALFDYDRTRDSCLPSQGLSFSY
PSD-95_dlgh4            ALFDYDKTKDCGFLSQALSFHF
                        *:**** .:*  : *:.* * .
```

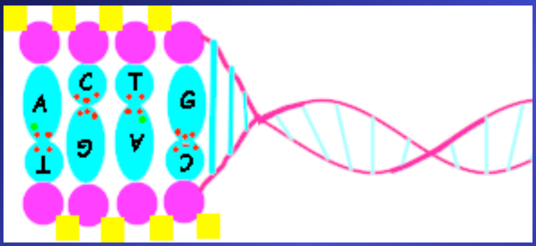
The consensus character is the one that minimises the distance between it and all the other characters in the column

Gene and Protein Prediction

Gene prediction

- What is a gene?
  - Simple definition: A stretch of DNA that encodes a protein and includes the regulatory sequences required for temporal and spatial control of gene transcription.
- Characteristics of genes.
  - What genetic features can we use to recognise a gene?

DNA structure



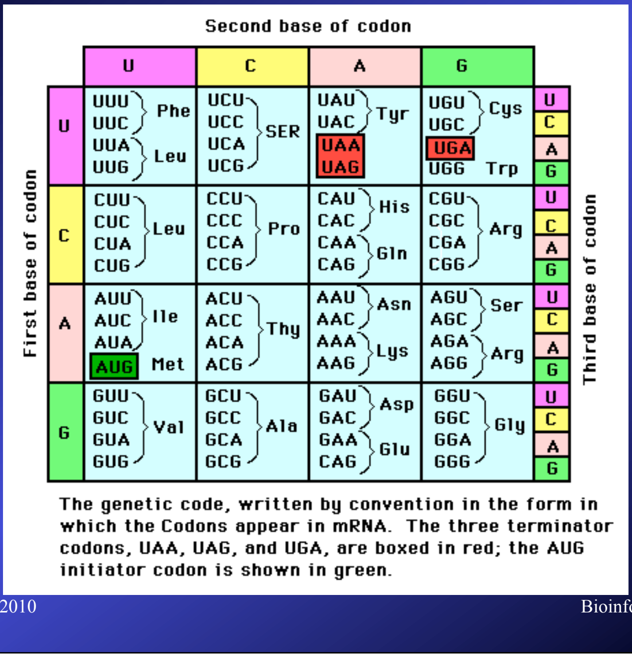
Bases: A,C,G and T

Chemically, A can only pair with T and G with C

Two strands, 5’ and 3’ Genes are encoded along one side of the DNA molecule. The 5’ end being at the left hand side of the gene.

Codons and ORFs

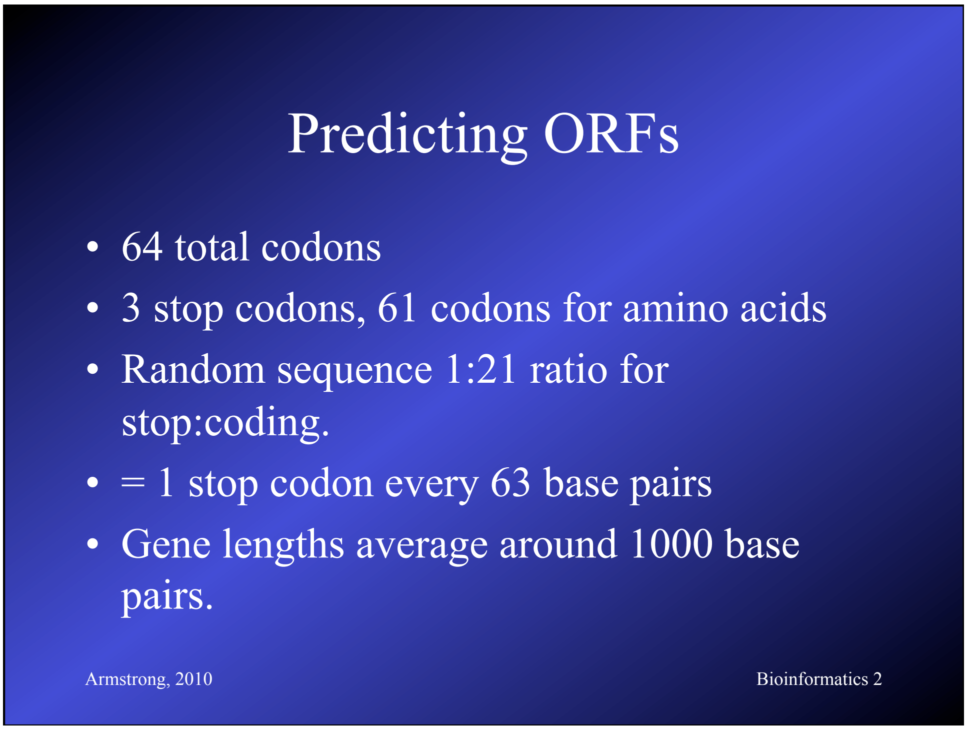
- Three bases that encode an amino acid or stop site.
- A run of valid codons is an Open Reading Frame.
- An ORF usually starts with a Met
- Ends with a nonsense or stop codon.



Second base of codon

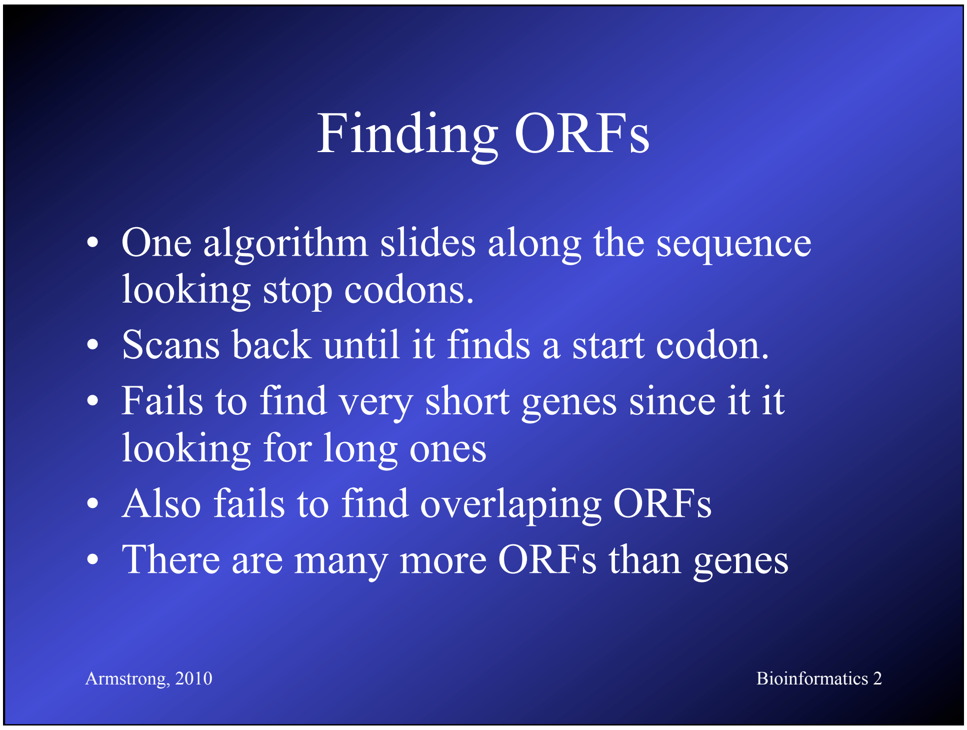
First base of codon	U	C	A	G	Third base of codon
U	UUU Phe	UCU SER	UAU Tyr	UGU Cys	U
	UUC	UCC	UAC	UGC	C
	UUA Leu	UCA	UAA	UGA	A
	UUG	UCG	UAG	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Gln	CGA	A
	CUG	CCG	CAG	CGG	G
A	AUU Ile	ACU Thy	AAU Asn	AGU Ser	U
	AUC	ACC	AAC	AGC	C
	AUA	ACA	AAA Lys	AGA Arg	A
	AUG Met	ACG	AAG	AGG	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA Glu	GGA	A
	GUG	GCG	CAG	GGG	G

The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.



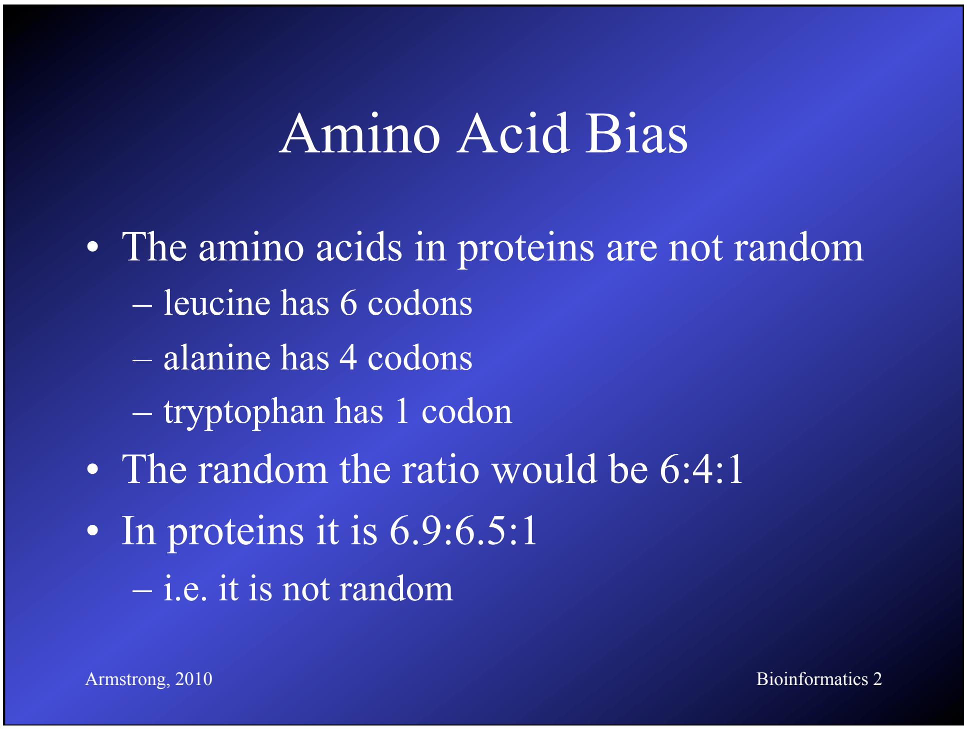
Predicting ORFs

- 64 total codons
- 3 stop codons, 61 codons for amino acids
- Random sequence 1:21 ratio for stop:coding.
- = 1 stop codon every 63 base pairs
- Gene lengths average around 1000 base pairs.



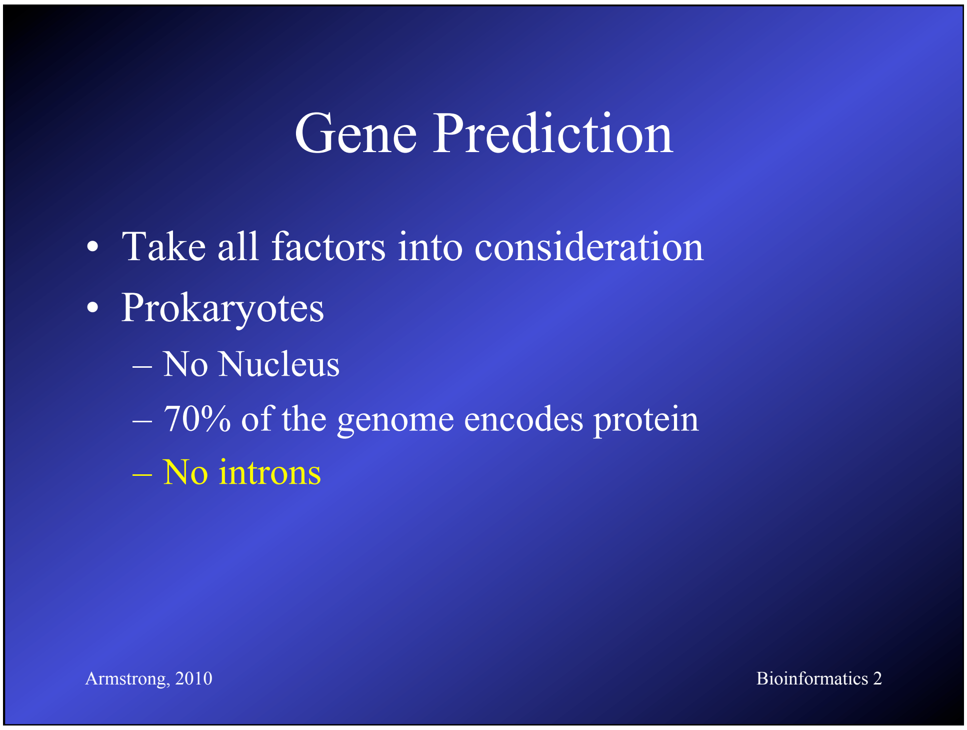
Finding ORFs

- One algorithm slides along the sequence looking stop codons.
- Scans back until it finds a start codon.
- Fails to find very short genes since it it looking for long ones
- Also fails to find overlaping ORFs
- There are many more ORFs than genes



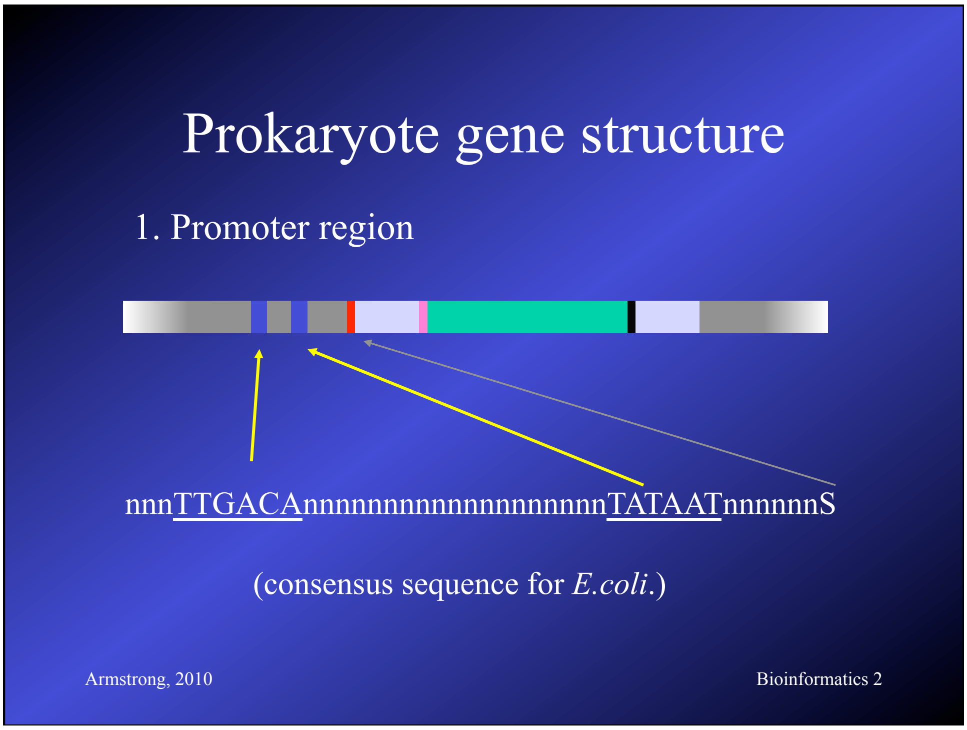
Amino Acid Bias

- The amino acids in proteins are not random
  - leucine has 6 codons
  - alanine has 4 codons
  - tryptophan has 1 codon
- The random the ratio would be 6:4:1
- In proteins it is 6.9:6.5:1
  - i.e. it is not random



Gene Prediction

- Take all factors into consideration
- Prokaryotes
  - No Nucleus
  - 70% of the genome encodes protein
  - No introns

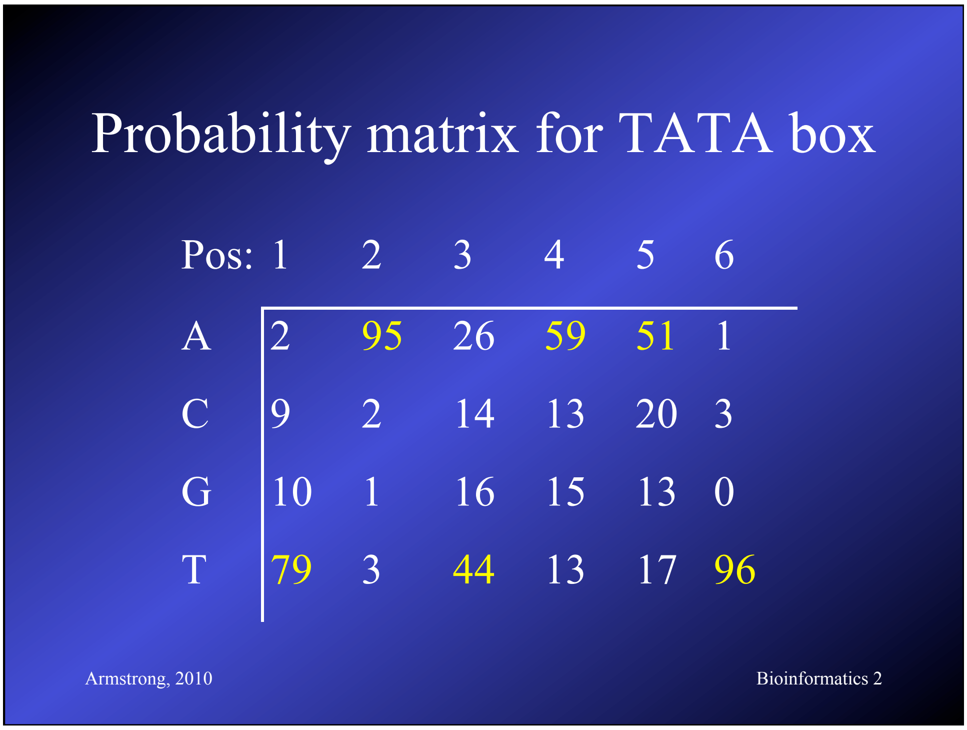


Prokaryote gene structure

1. Promoter region

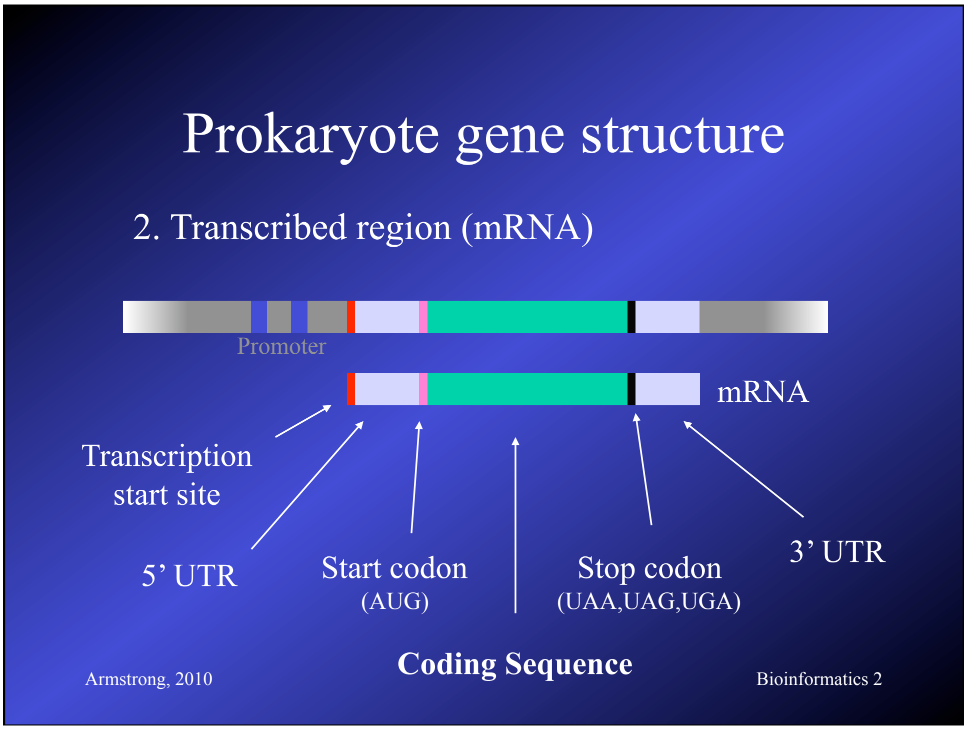
nnnTTGACAnnnnnnnnnnnnnnnnnnTATAATnnnnnnS

(consensus sequence for *E.coli*.)



Probability matrix for TATA box

Pos:	1	2	3	4	5	6
A	2	95	26	59	51	1
C	9	2	14	13	20	3
G	10	1	16	15	13	0
T	79	3	44	13	17	96



Prokaryote gene structure

2. Transcribed region (mRNA)

Promoter

mRNA

Transcription start site

5' UTR

Start codon (AUG)

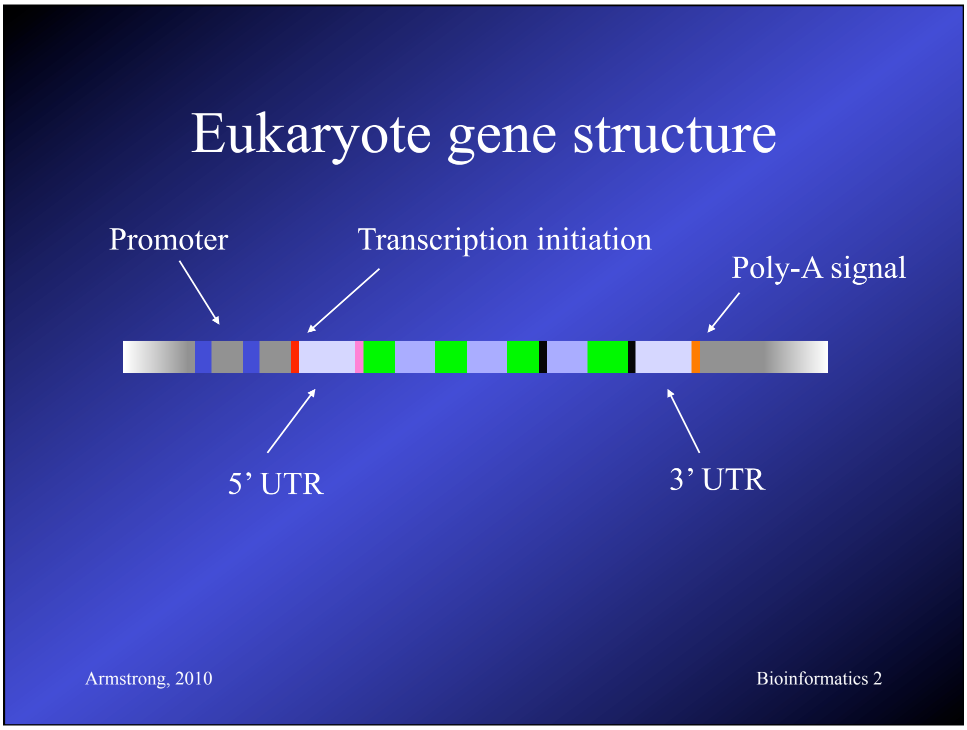
Stop codon (UAA,UAG,UGA)

3' UTR

Coding Sequence

Armstrong, 2010

Bioinformatics 2



Eukaryote gene structure

Promoter

Transcription initiation

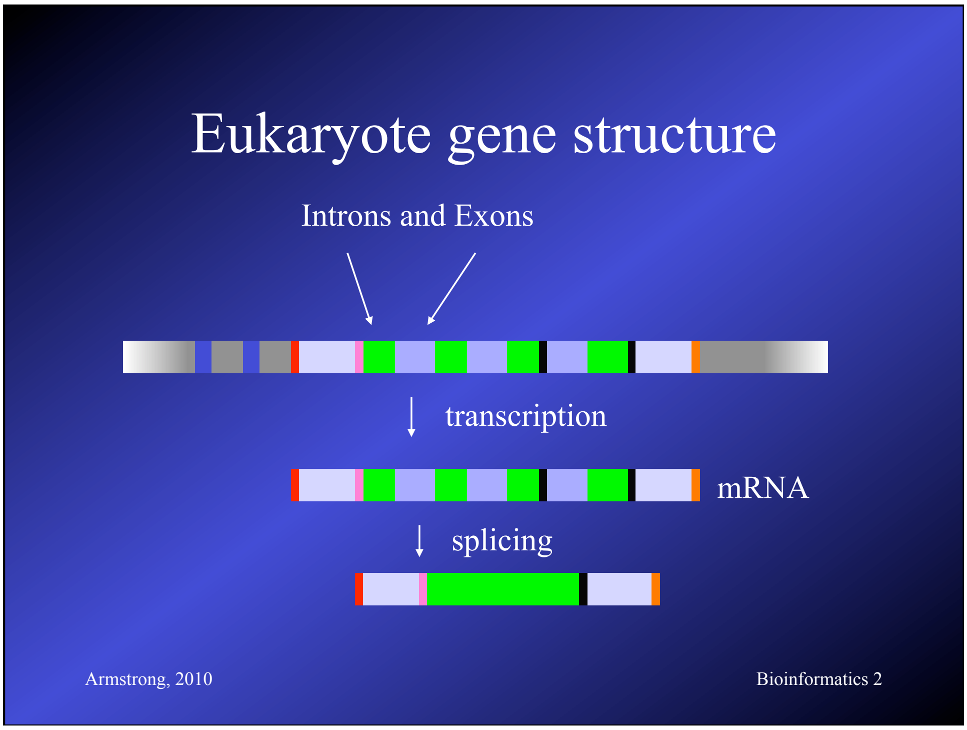
Poly-A signal

5' UTR

3' UTR

Armstrong, 2010

Bioinformatics 2



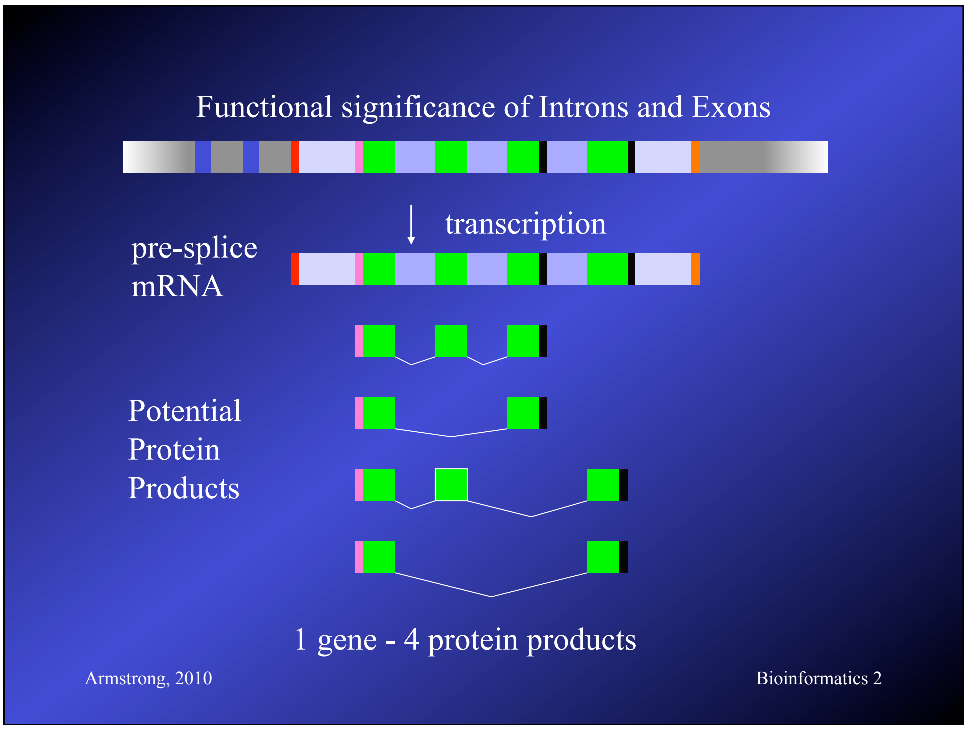
Eukaryote gene structure

Introns and Exons

transcription

mRNA

splicing



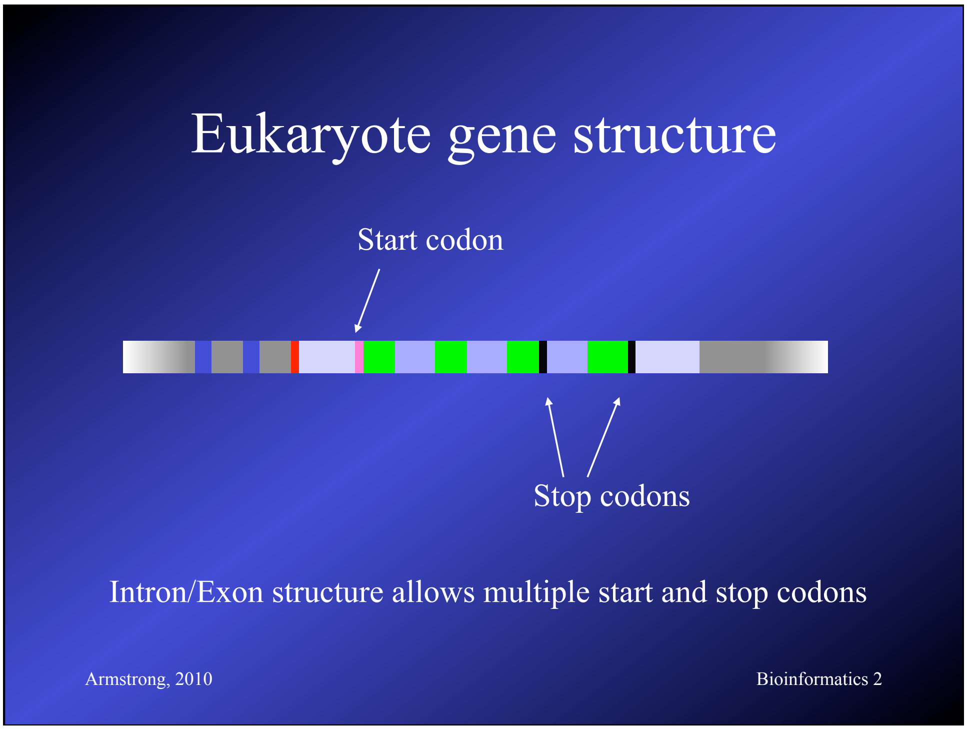
Functional significance of Introns and Exons

transcription

pre-splice mRNA

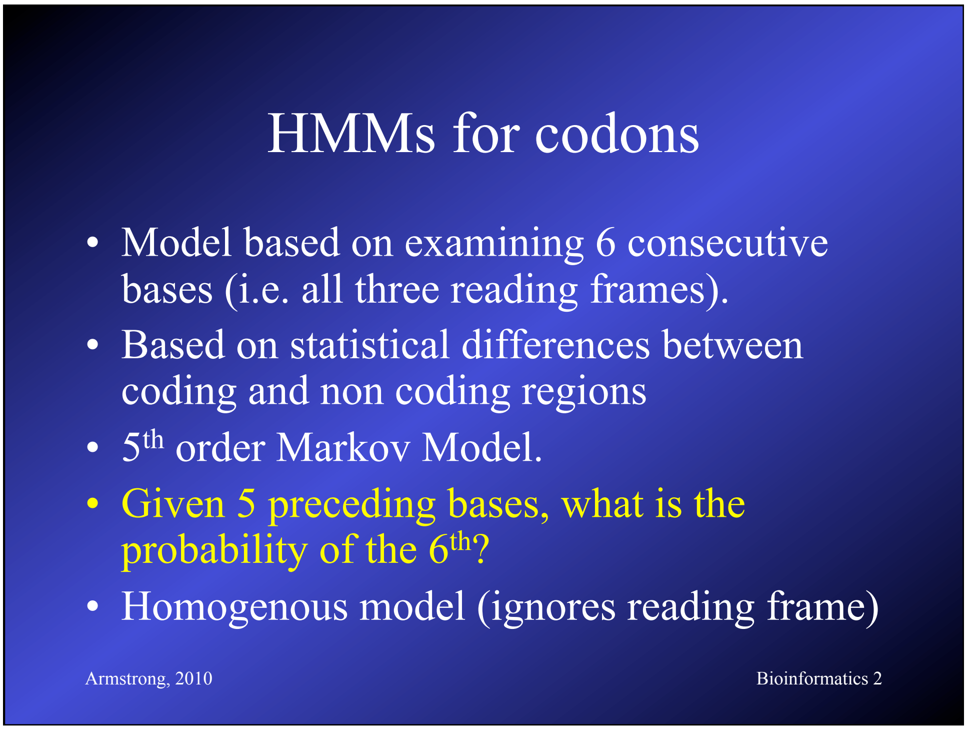
Potential Protein Products

1 gene - 4 protein products



Eukaryote gene structure

Intron/Exon structure allows multiple start and stop codons



HMMs for codons

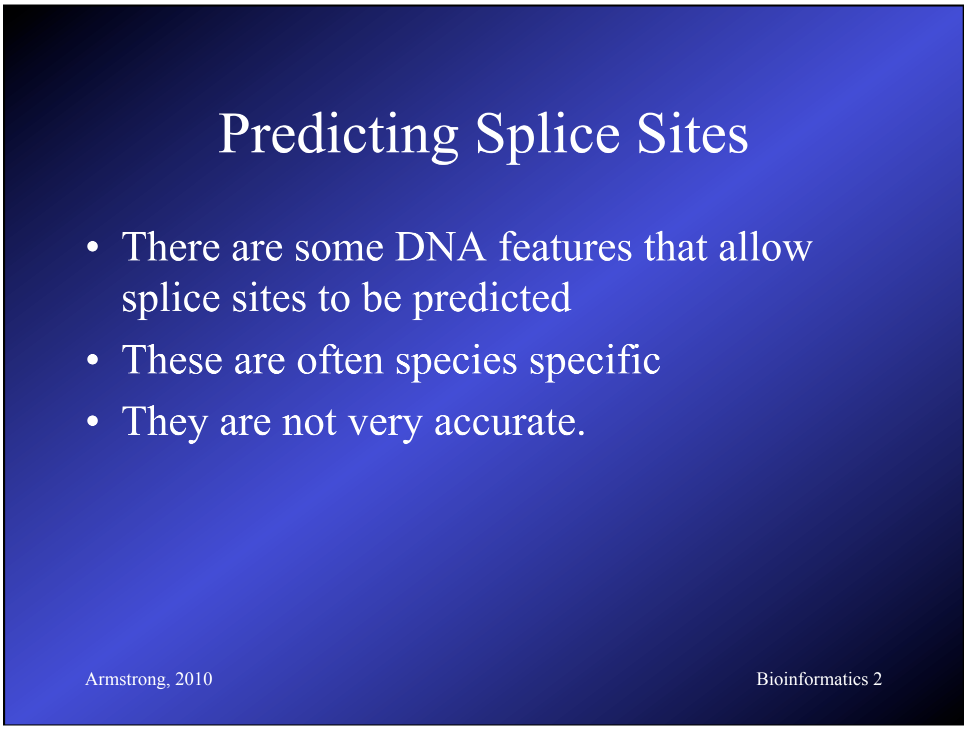
- Model based on examining 6 consecutive bases (i.e. all three reading frames).
- Based on statistical differences between coding and non coding regions
- 5th order Markov Model.
- Given 5 preceding bases, what is the probability of the 6th?
- Homogenous model (ignores reading frame)

HMMs for codons

- Homogenous models have two tables, one for coding, one non coding.
- Each table is has 4096 entries for the potential 6 base pair sequences
- Non-homogenous models have three tables for possible reading frames
- Short exons cause these models problems
- Hard to detect splice sites

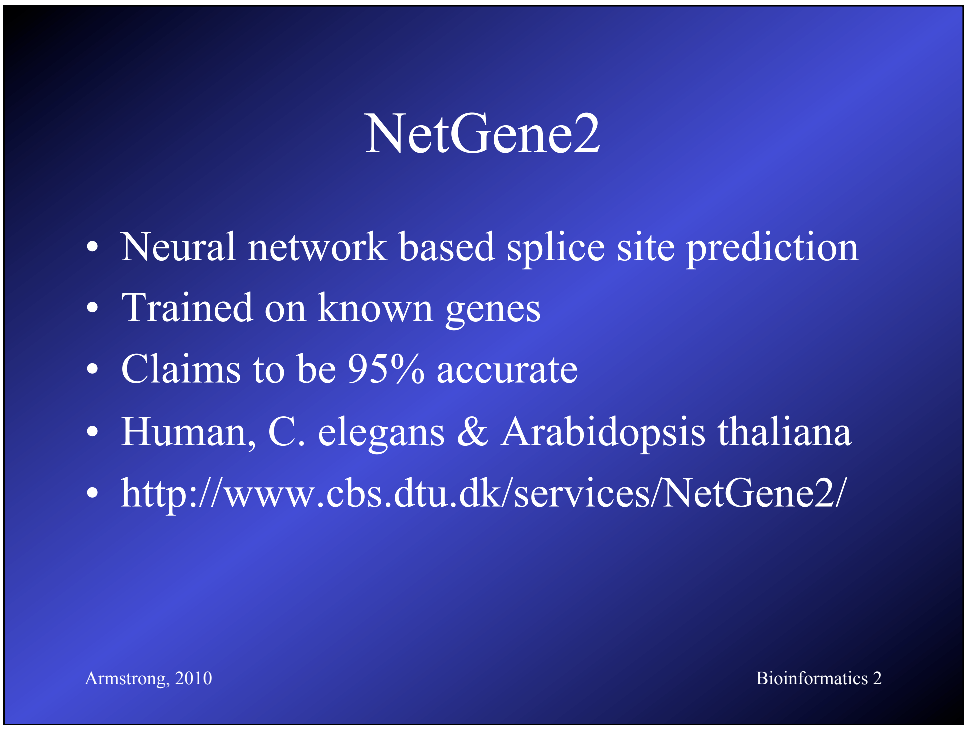
Glimmer

- Uses non-homogenous HMMs to predict prokaryote gene sequences
- Identifies ORFs
- Trains itself on a prokaryote genome using ORFs over 500 bp
- http://www.cs.jhu.edu/labs/compbio/glimmer.html



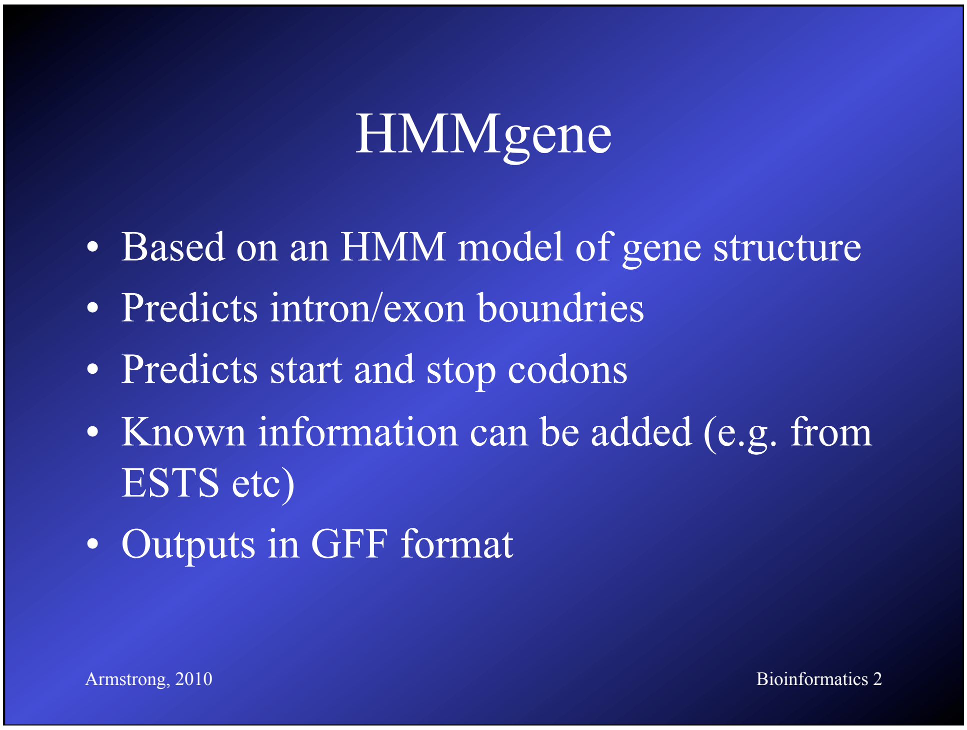
Predicting Splice Sites

- There are some DNA features that allow splice sites to be predicted
- These are often species specific
- They are not very accurate.



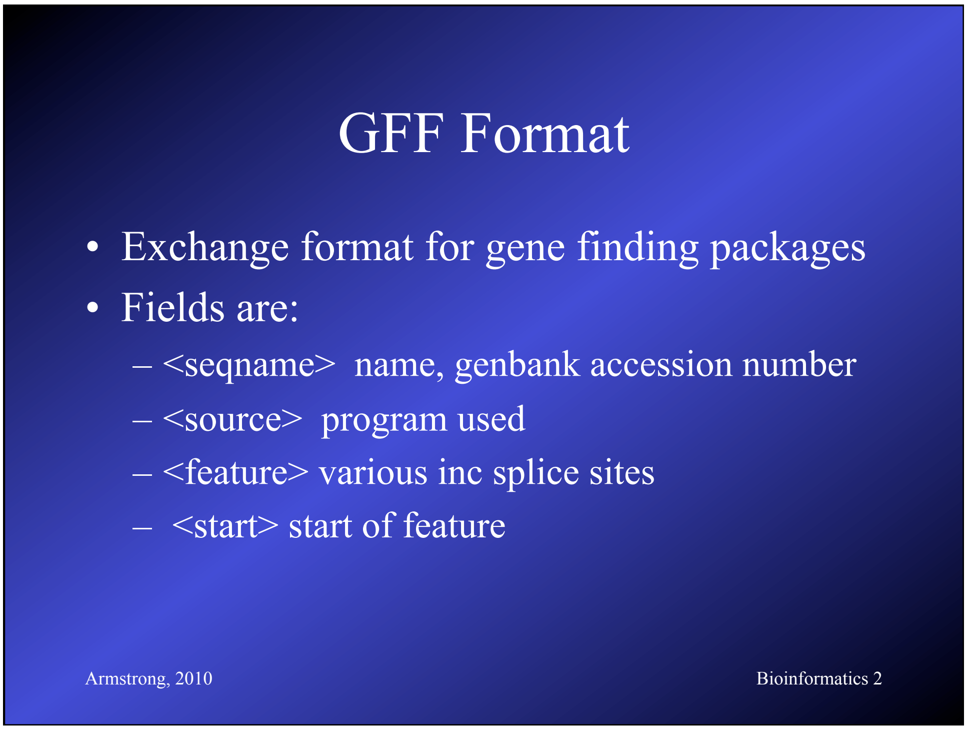
NetGene2

- Neural network based splice site prediction
- Trained on known genes
- Claims to be 95% accurate
- Human, C. elegans & Arabidopsis thaliana
- http://www.cbs.dtu.dk/services/NetGene2/



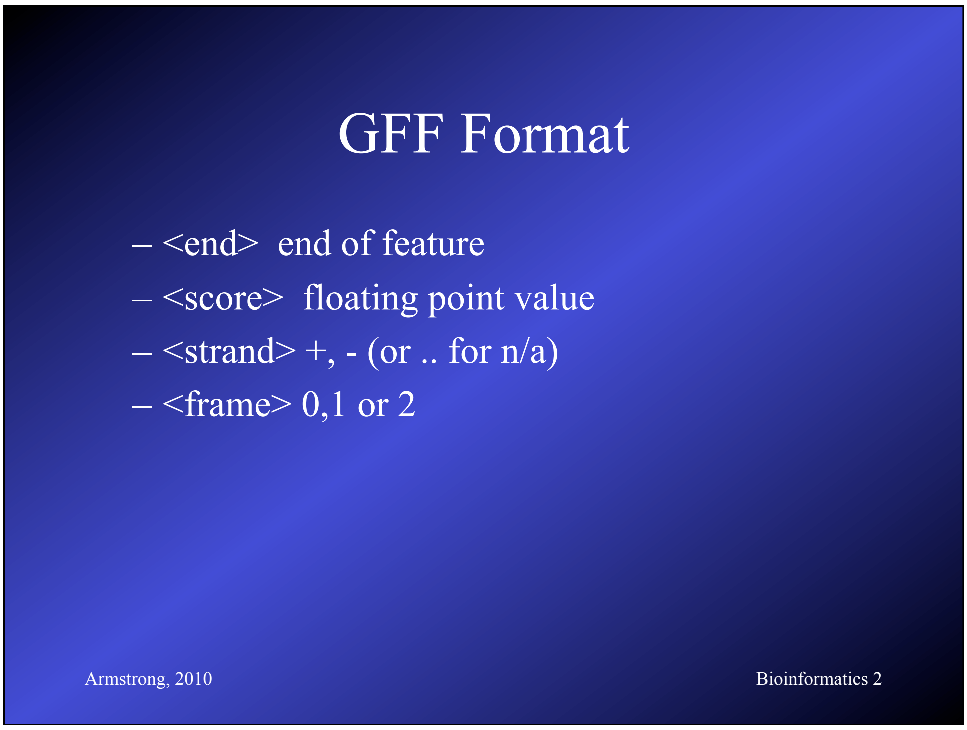
HMMgene

- Based on an HMM model of gene structure
- Predicts intron/exon boundries
- Predicts start and stop codons
- Known information can be added (e.g. from ESTS etc)
- Outputs in GFF format



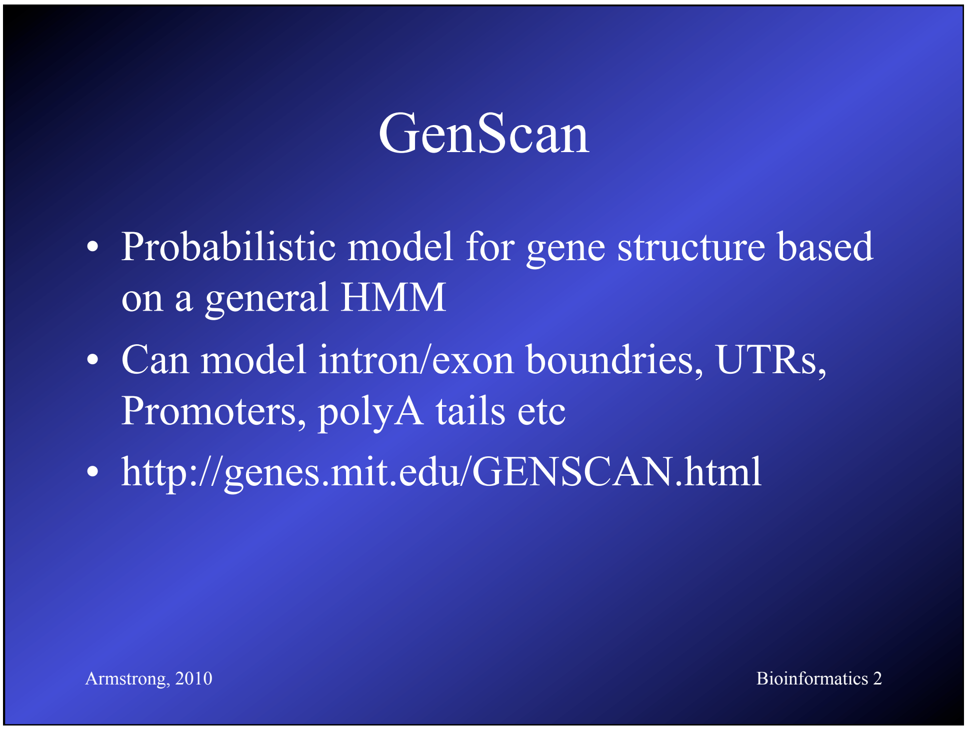
GFF Format

- Exchange format for gene finding packages
- Fields are:
  - <seqname> name, genbank accession number
  - <source> program used
  - <feature> various inc splice sites
  - <start> start of feature



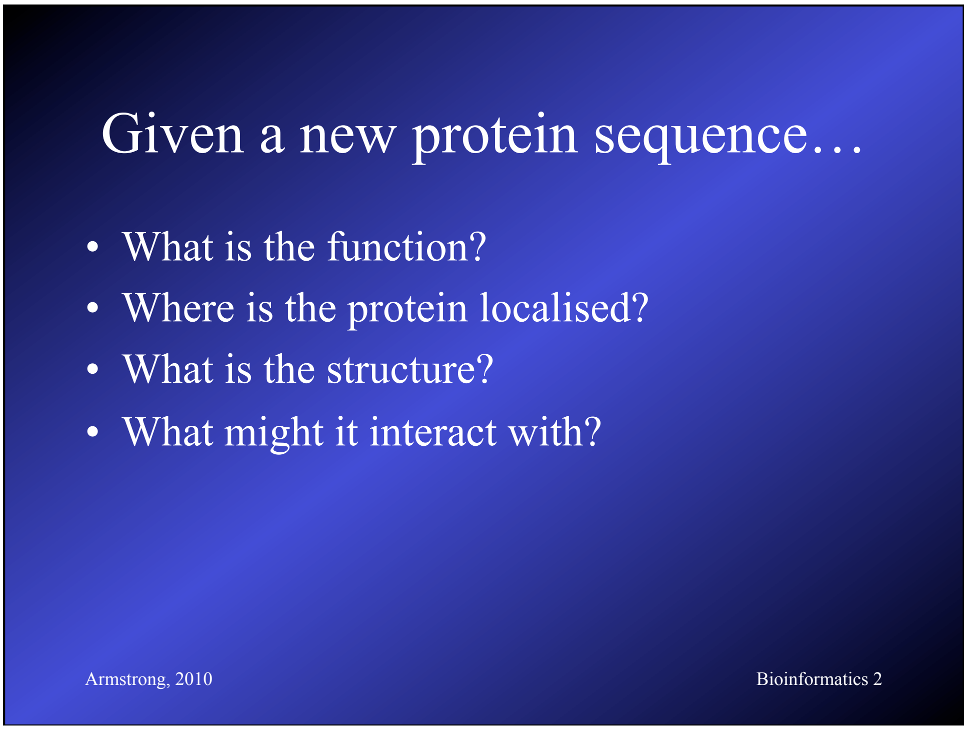
GFF Format

- <end> end of feature
- <score> floating point value
- <strand> +, - (or .. for n/a)
- <frame> 0,1 or 2



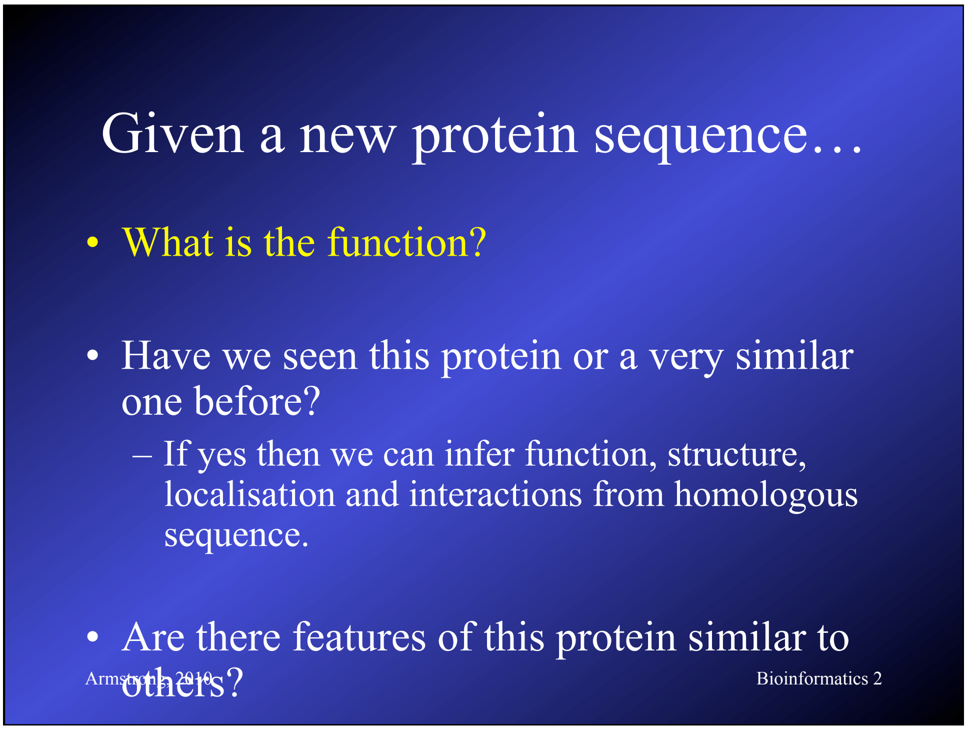
GenScan

- Probabilistic model for gene structure based on a general HMM
- Can model intron/exon boundries, UTRs, Promoters, polyA tails etc
- http://genes.mit.edu/GENSCAN.html



Given a new protein sequence…

- What is the function?
- Where is the protein localised?
- What is the structure?
- What might it interact with?

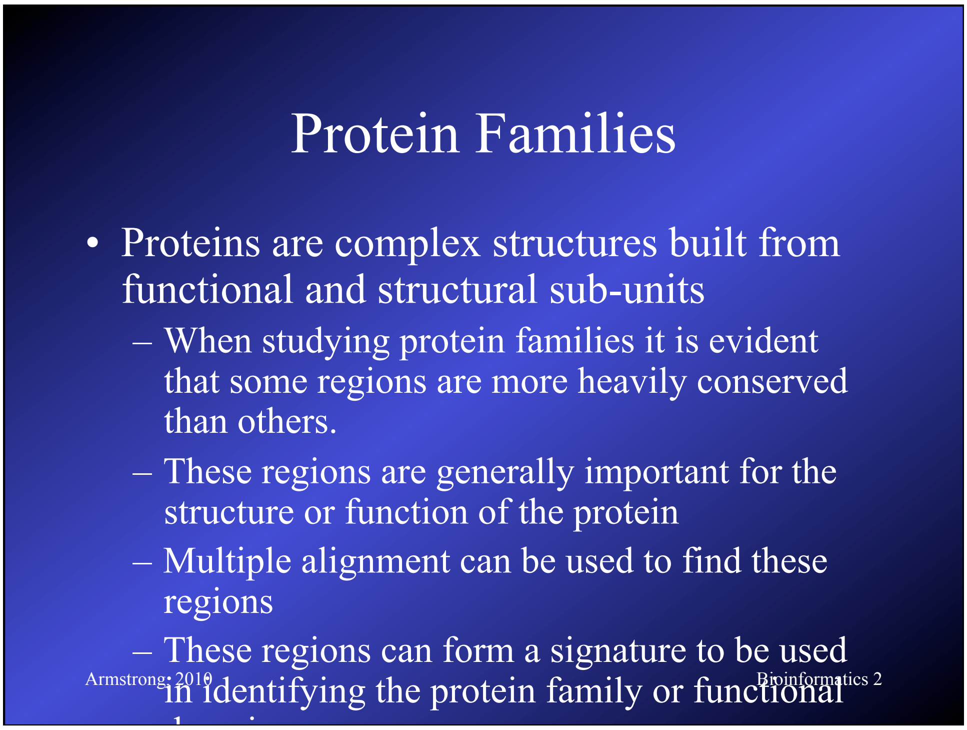


Given a new protein sequence…

- What is the function?

- Have we seen this protein or a very similar one before?
  - If yes then we can infer function, structure, localisation and interactions from homologous sequence.

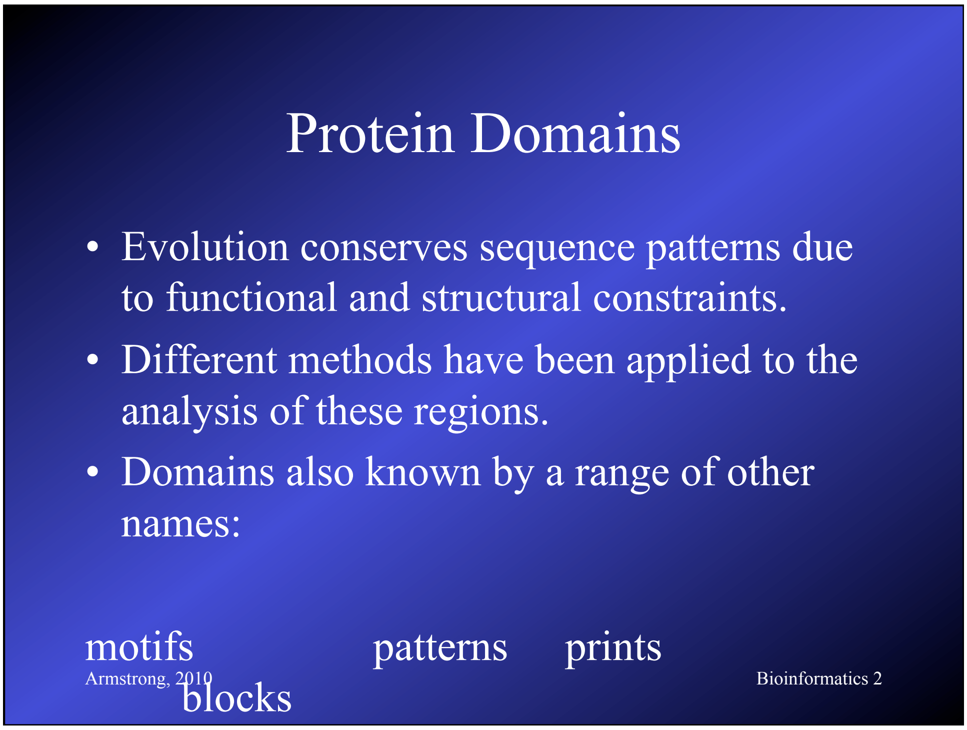
- Are there features of this protein similar to others?



Protein Families

- Proteins are complex structures built from functional and structural sub-units
  - When studying protein families it is evident that some regions are more heavily conserved than others.
  - These regions are generally important for the structure or function of the protein
  - Multiple alignment can be used to find these regions
  - These regions can form a signature to be used in identifying the protein family or functional

Armstrong, 2010 Bioinformatics 2



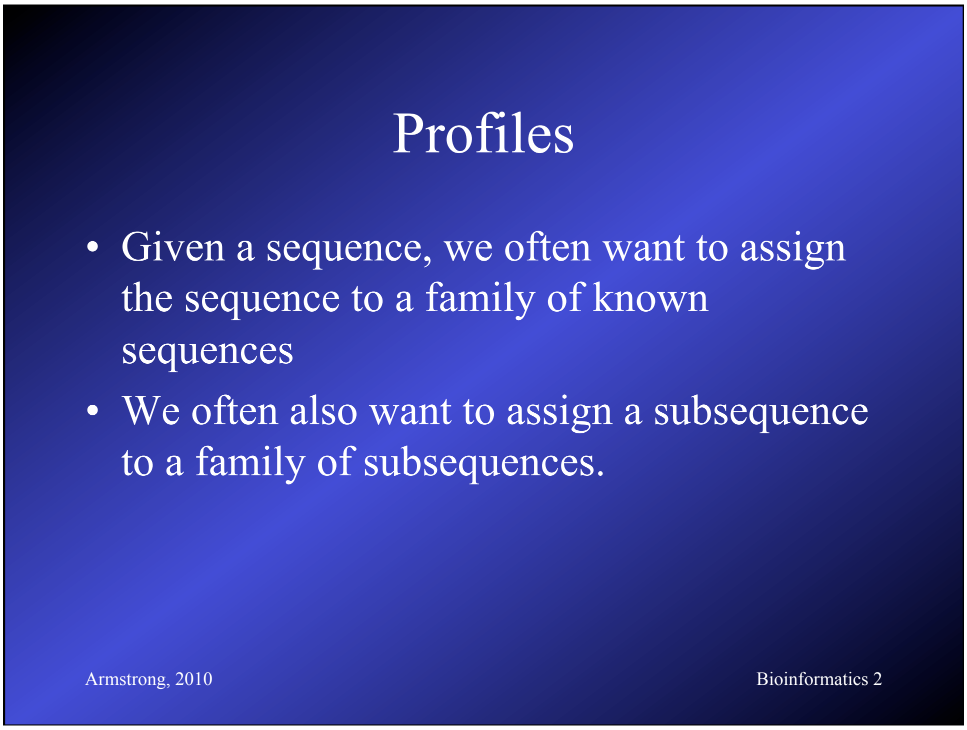
Protein Domains

- Evolution conserves sequence patterns due to functional and structural constraints.
- Different methods have been applied to the analysis of these regions.
- Domains also known by a range of other names:

motifs patterns prints

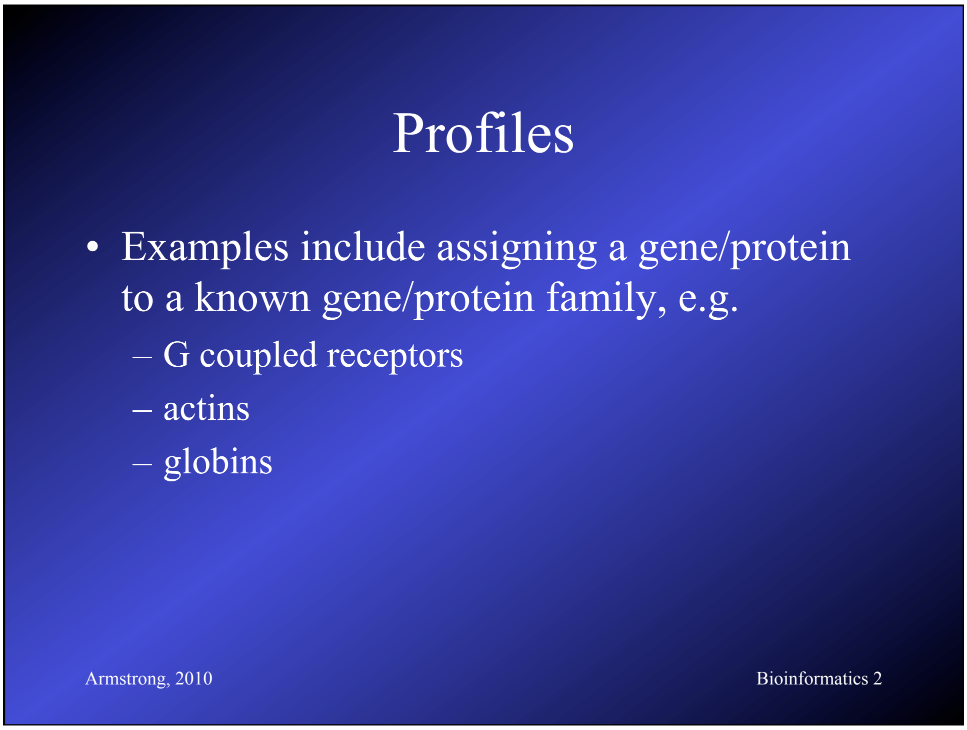
blocks

Armstrong, 2010 Bioinformatics 2



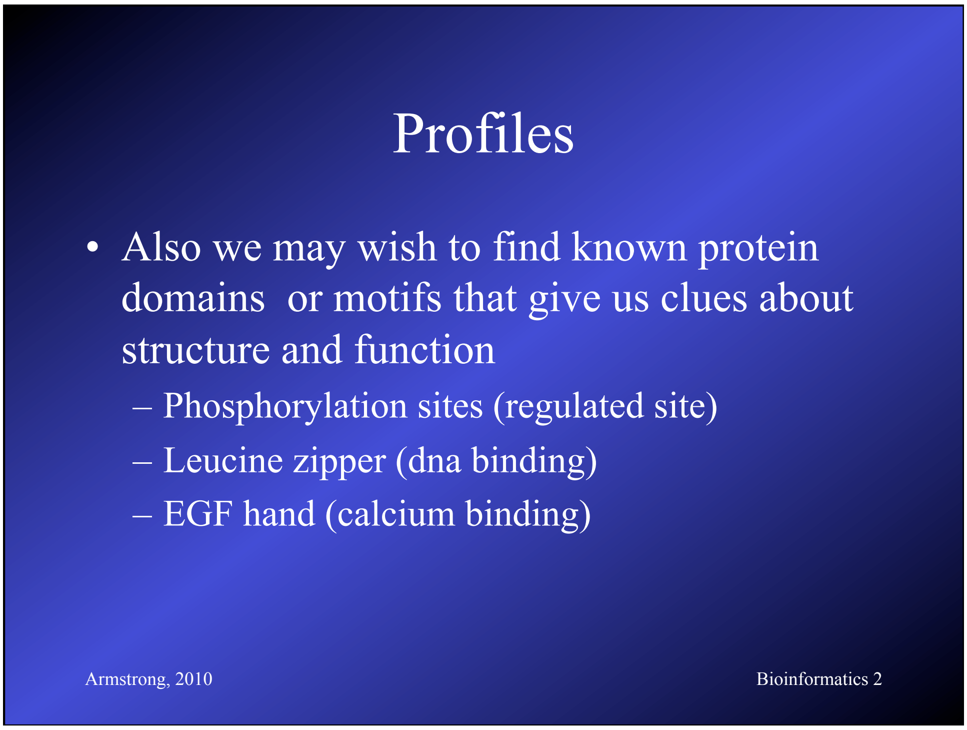
Profiles

- Given a sequence, we often want to assign the sequence to a family of known sequences
- We often also want to assign a subsequence to a family of subsequences.



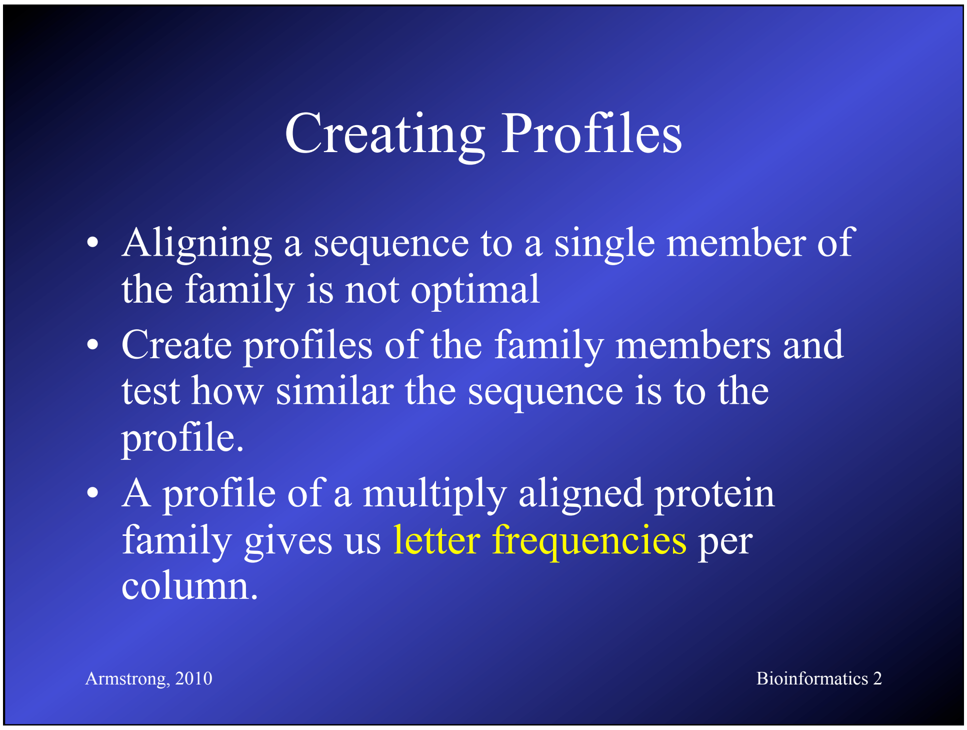
Profiles

- Examples include assigning a gene/protein to a known gene/protein family, e.g.
  - G coupled receptors
  - actins
  - globins



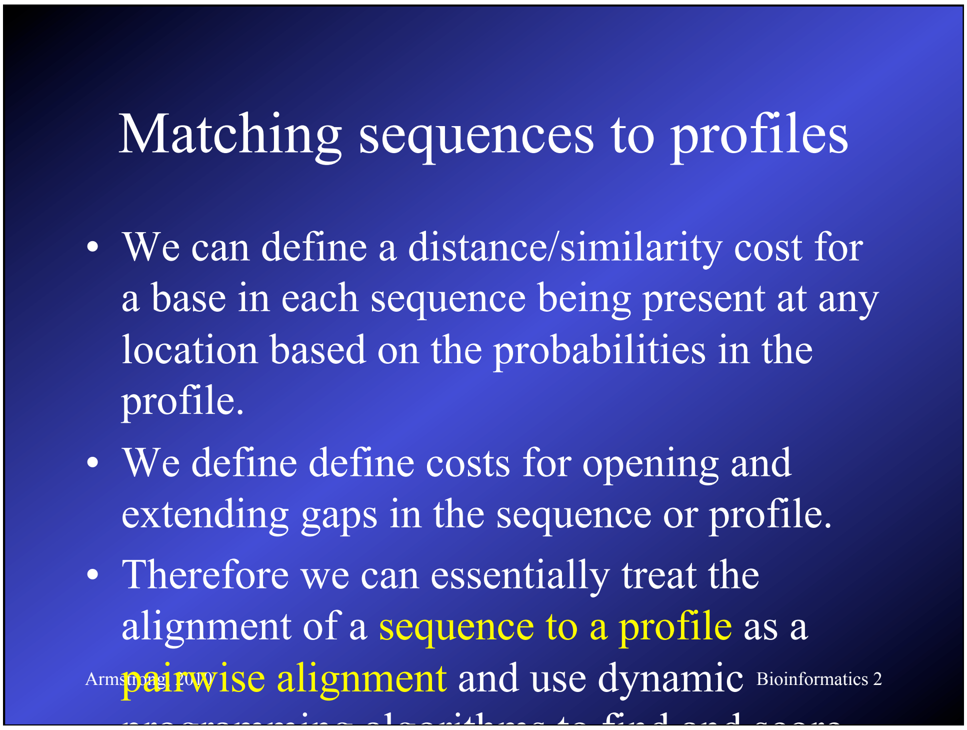
Profiles

- Also we may wish to find known protein domains or motifs that give us clues about structure and function
  - Phosphorylation sites (regulated site)
  - Leucine zipper (dna binding)
  - EGF hand (calcium binding)



Creating Profiles

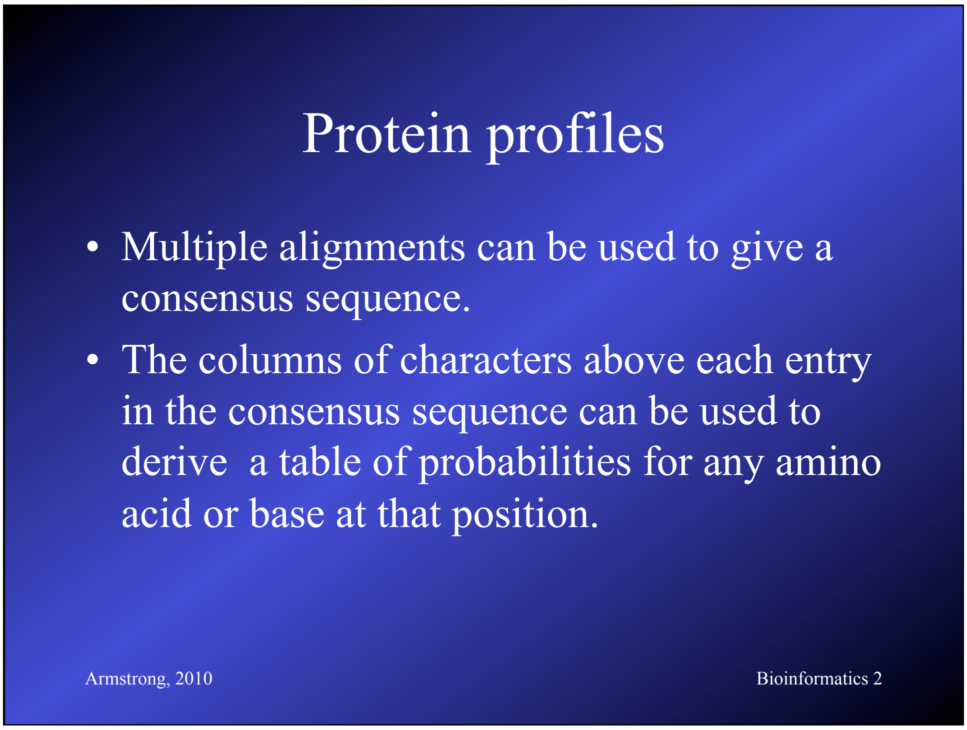
- Aligning a sequence to a single member of the family is not optimal
- Create profiles of the family members and test how similar the sequence is to the profile.
- A profile of a multiply aligned protein family gives us letter frequencies per column.



Matching sequences to profiles

- We can define a distance/similarity cost for a base in each sequence being present at any location based on the probabilities in the profile.
- We define define costs for opening and extending gaps in the sequence or profile.
- Therefore we can essentially treat the alignment of a sequence to a profile as a pairwise alignment and use dynamic programming algorithms to find and score

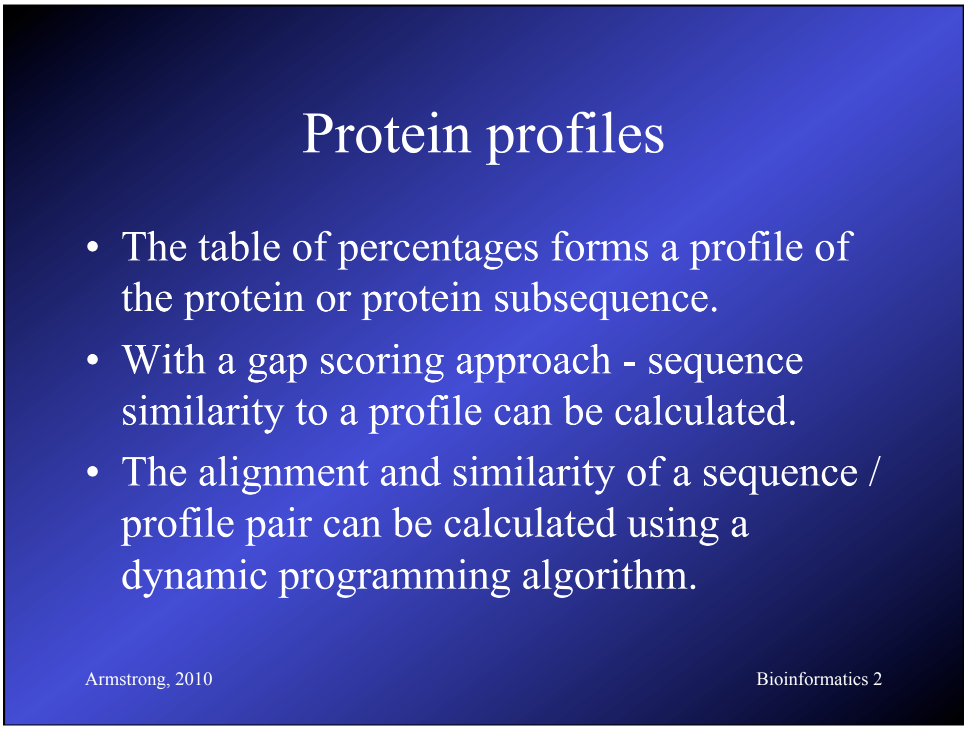
Armstrong, 2010 Bioinformatics 2



Protein profiles

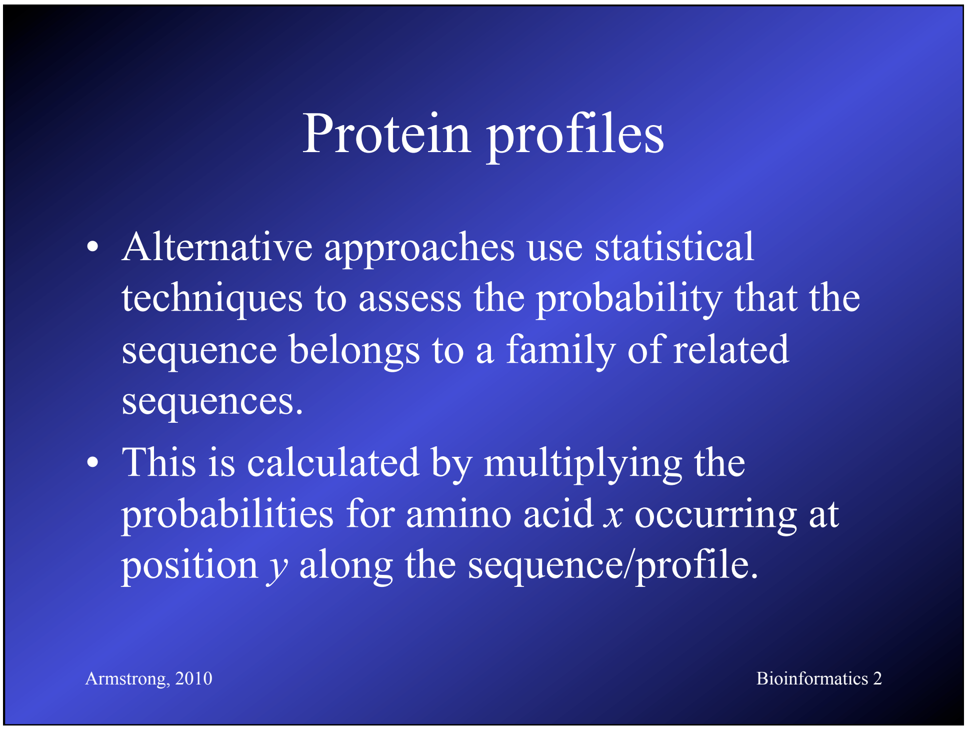
- Multiple alignments can be used to give a consensus sequence.
- The columns of characters above each entry in the consensus sequence can be used to derive a table of probabilities for any amino acid or base at that position.

Armstrong, 2010 Bioinformatics 2



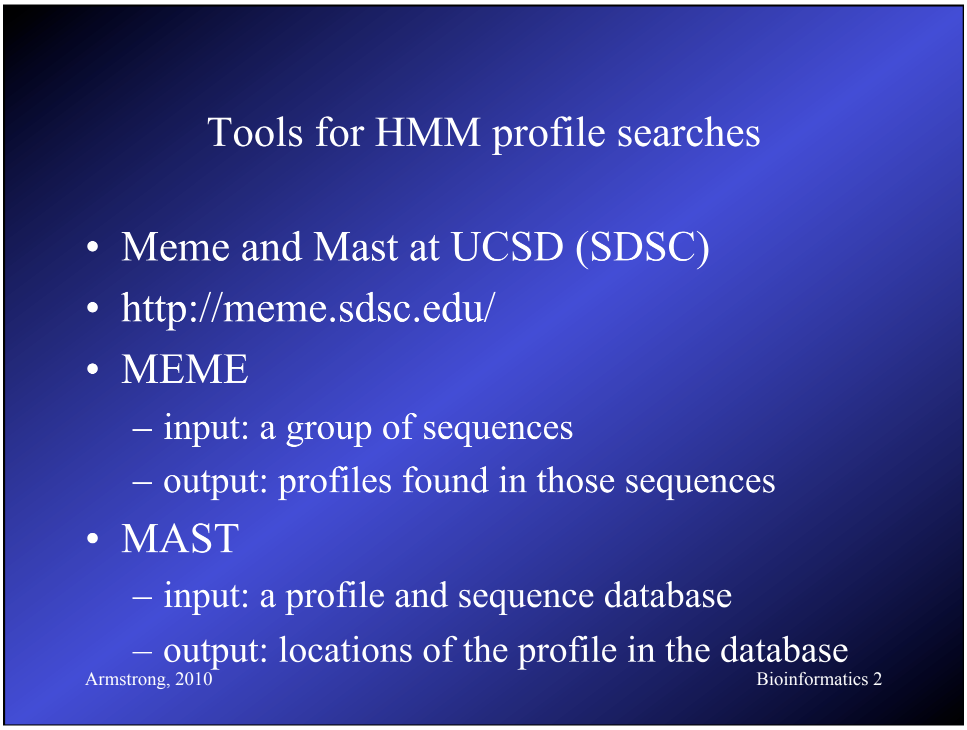
Protein profiles

- The table of percentages forms a profile of the protein or protein subsequence.
- With a gap scoring approach - sequence similarity to a profile can be calculated.
- The alignment and similarity of a sequence / profile pair can be calculated using a dynamic programming algorithm.



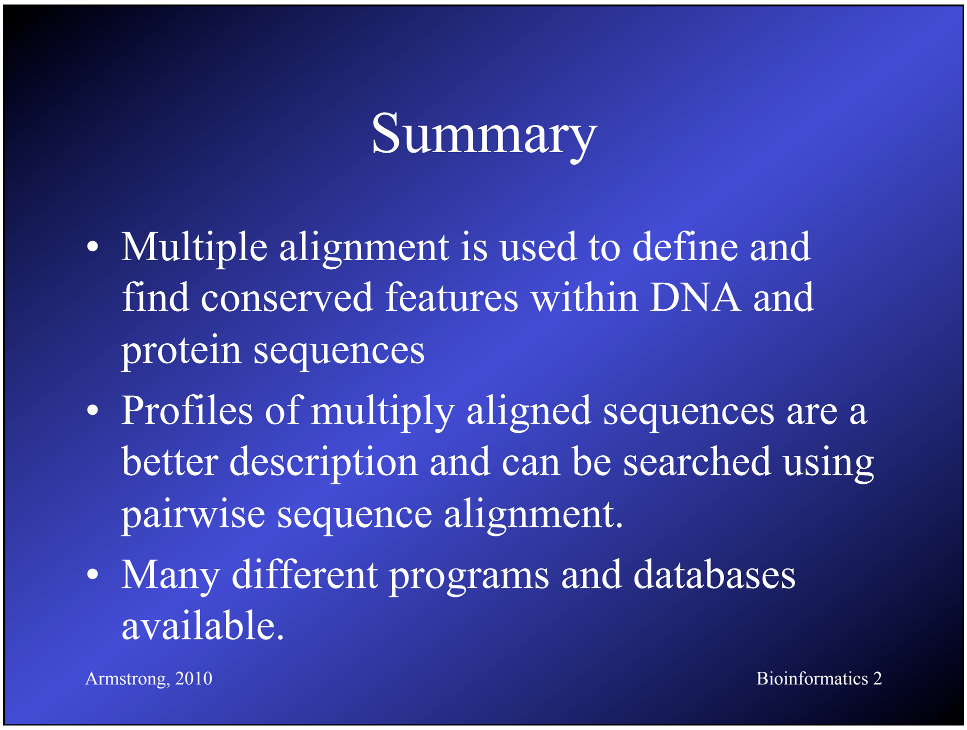
Protein profiles

- Alternative approaches use statistical techniques to assess the probability that the sequence belongs to a family of related sequences.
- This is calculated by multiplying the probabilities for amino acid *x* occurring at position *y* along the sequence/profile.



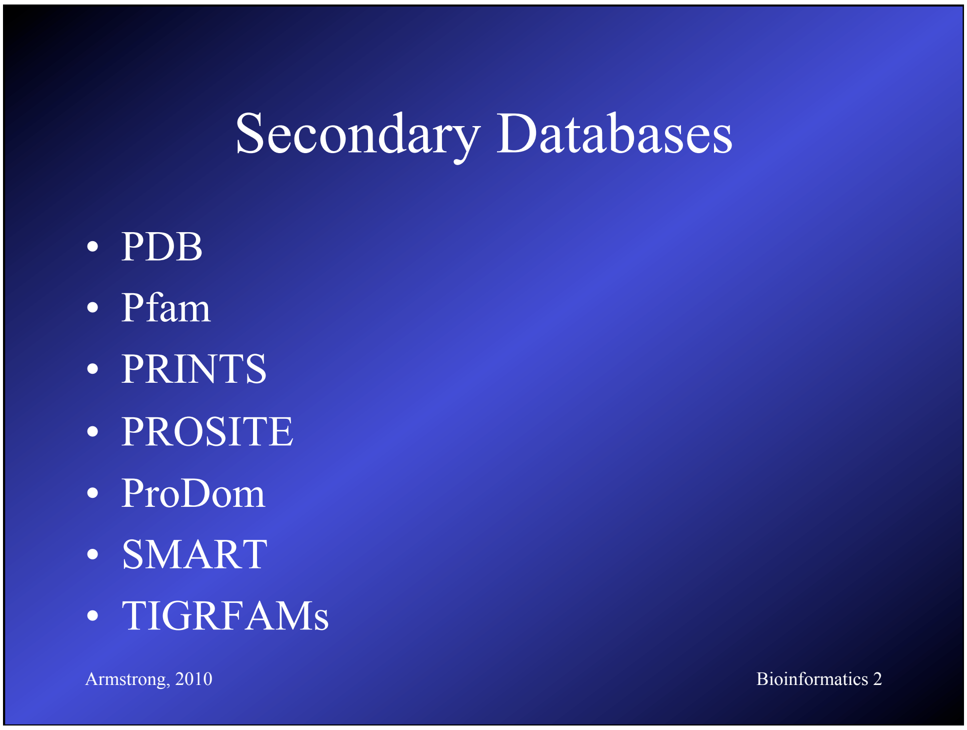
Tools for HMM profile searches

- Meme and Mast at UCSD (SDSC)
- http://meme.sdsc.edu/
- MEME
  - input: a group of sequences
  - output: profiles found in those sequences
- MAST
  - input: a profile and sequence database
  - output: locations of the profile in the database



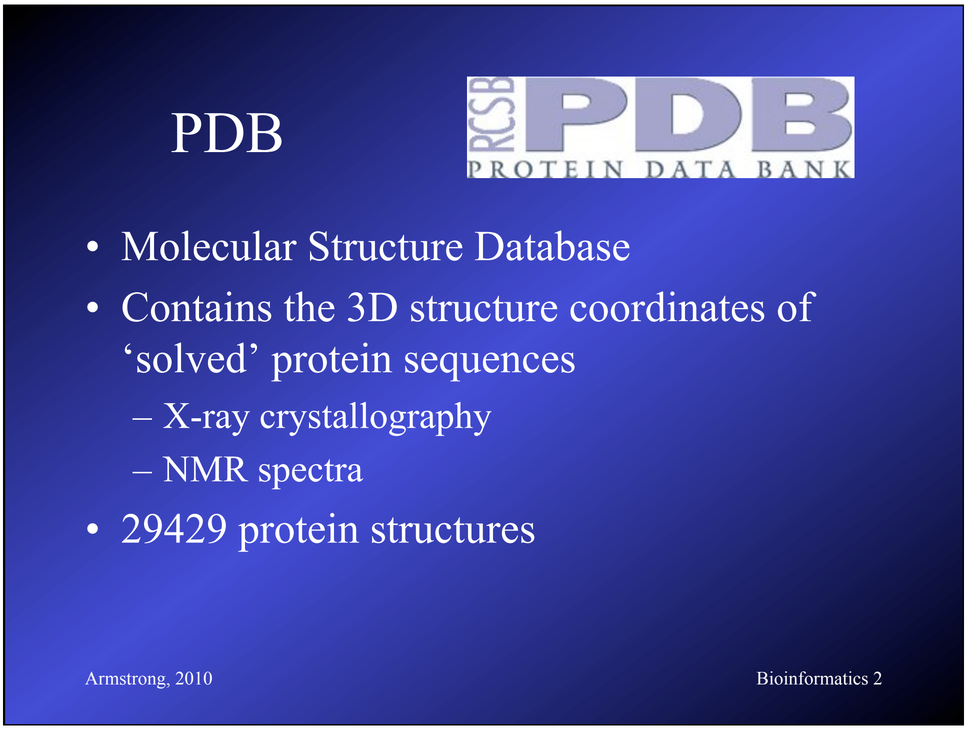
Summary

- Multiple alignment is used to define and find conserved features within DNA and protein sequences
- Profiles of multiply aligned sequences are a better description and can be searched using pairwise sequence alignment.
- Many different programs and databases available.



Secondary Databases

- PDB
- Pfam
- PRINTS
- PROSITE
- ProDom
- SMART
- TIGRFAMs



PDB

RCSB PDB PROTEIN DATA BANK

- Molecular Structure Database
- Contains the 3D structure coordinates of ‘solved’ protein sequences
  - X-ray crystallography
  - NMR spectra
- 29429 protein structures

Superfamily 1.65

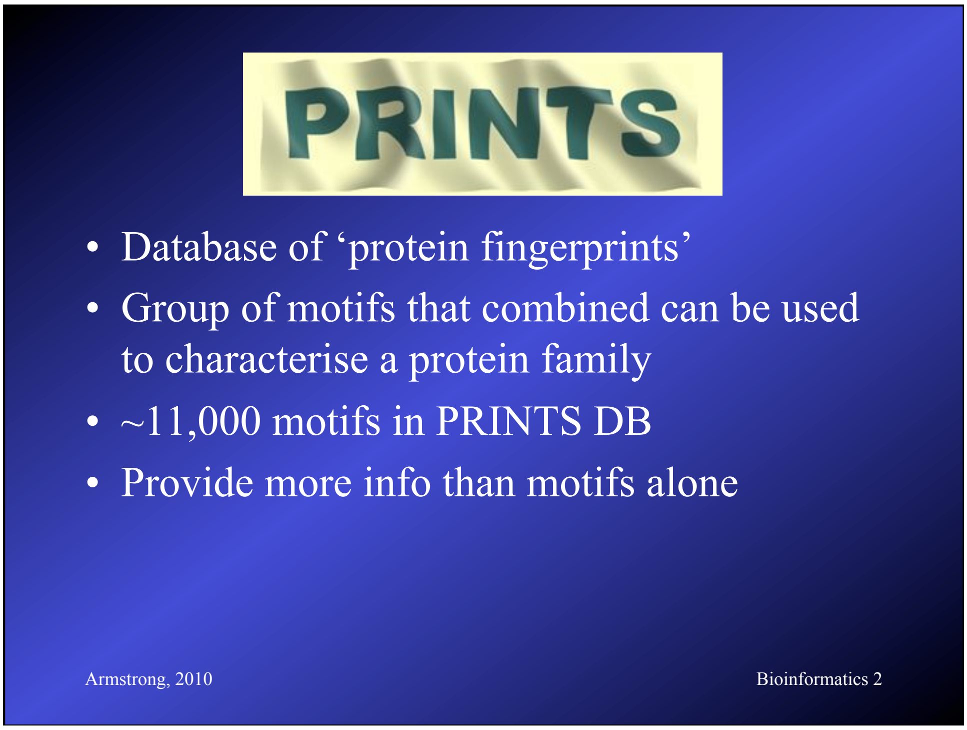
HMM library and genome assignments server

SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure, based on SCOP.

The SCOP database aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known (based on PDB)

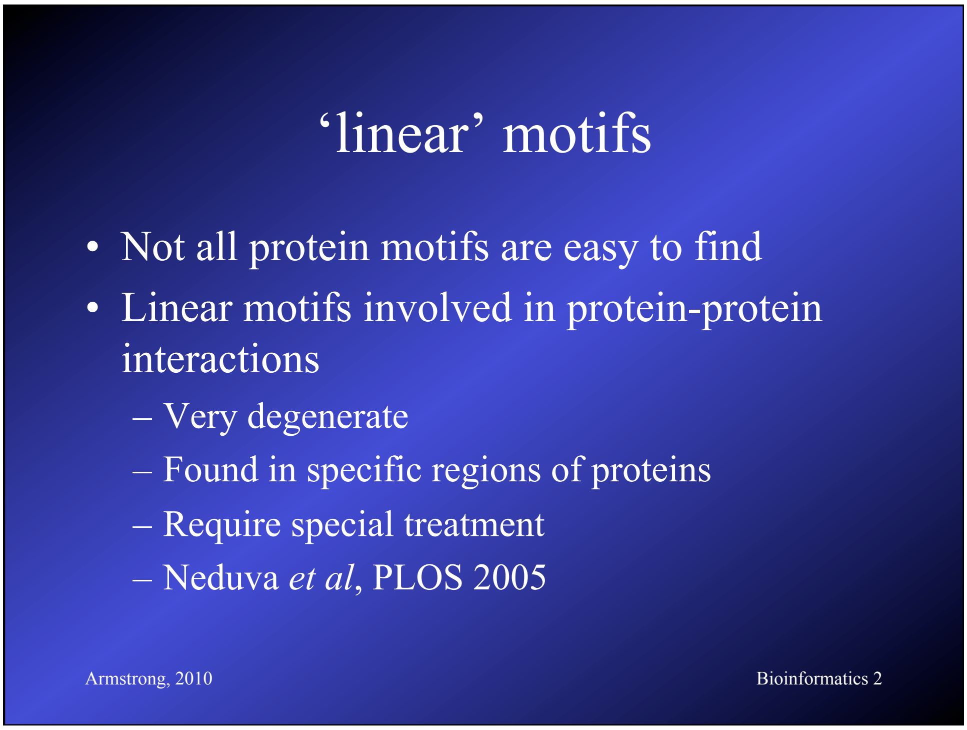
Pfam

- Database of protein domains
- Multiple sequence alignments and profile HMMs
- Entries also annotated
- Swiss-Prot DB all pre-searched
- New sequences can be searched as well.
  - 7973 entries in Pfam last update



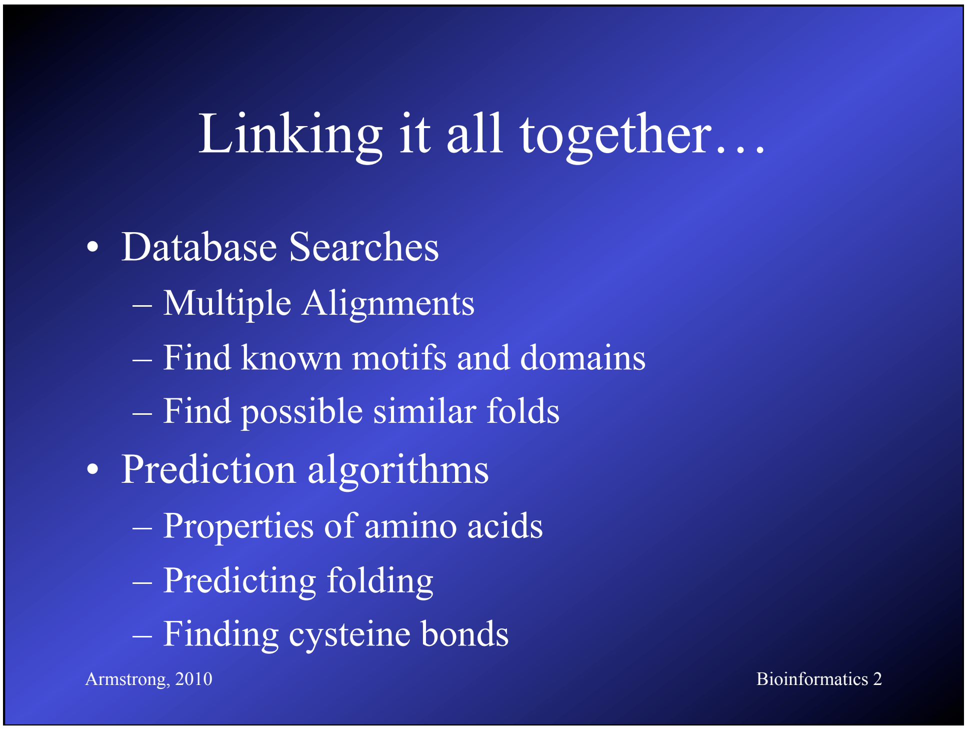
PRINTS

- Database of ‘protein fingerprints’
- Group of motifs that combined can be used to characterise a protein family
- ~11,000 motifs in PRINTS DB
- Provide more info than motifs alone



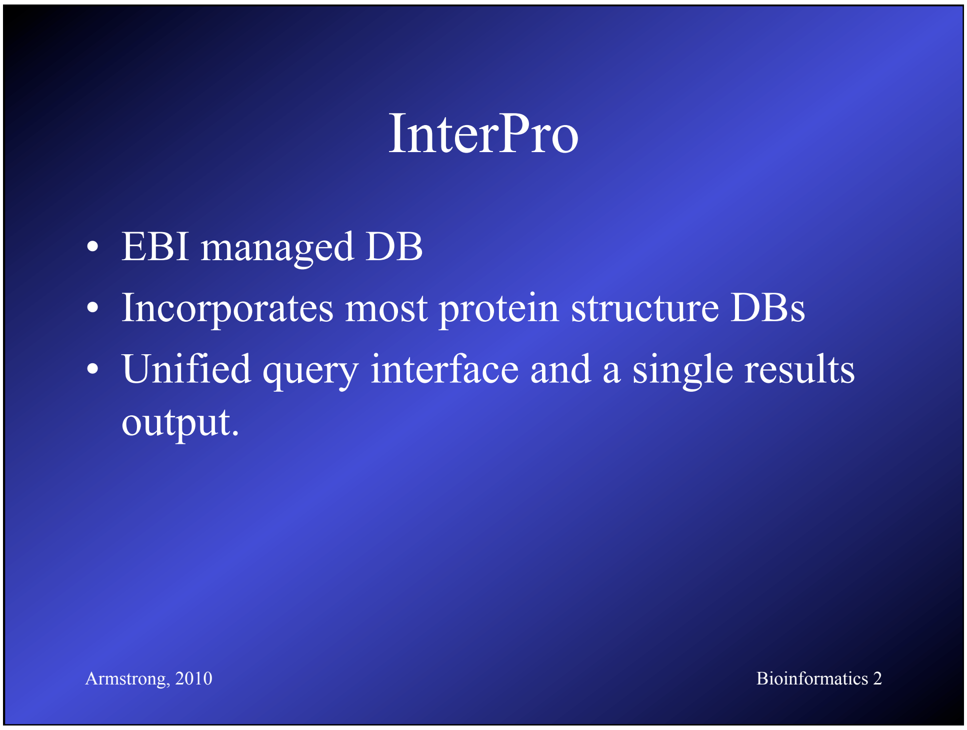
‘linear’ motifs

- Not all protein motifs are easy to find
- Linear motifs involved in protein-protein interactions
  - Very degenerate
  - Found in specific regions of proteins
  - Require special treatment
  - Neduva *et al*, PLOS 2005



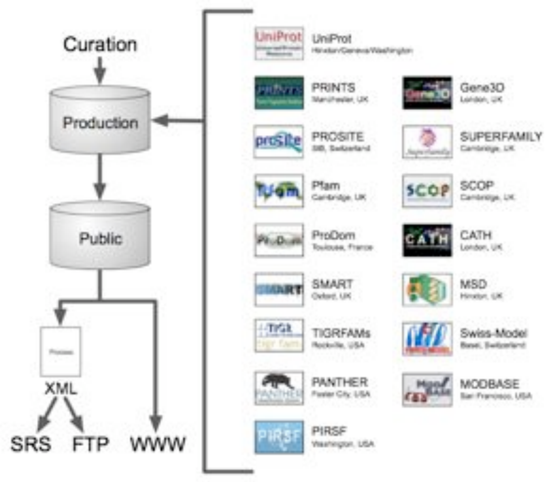
Linking it all together…

- Database Searches
  - Multiple Alignments
  - Find known motifs and domains
  - Find possible similar folds
- Prediction algorithms
  - Properties of amino acids
  - Predicting folding
  - Finding cysteine bonds



InterPro

- EBI managed DB
- Incorporates most protein structure DBs
- Unified query interface and a single results output.

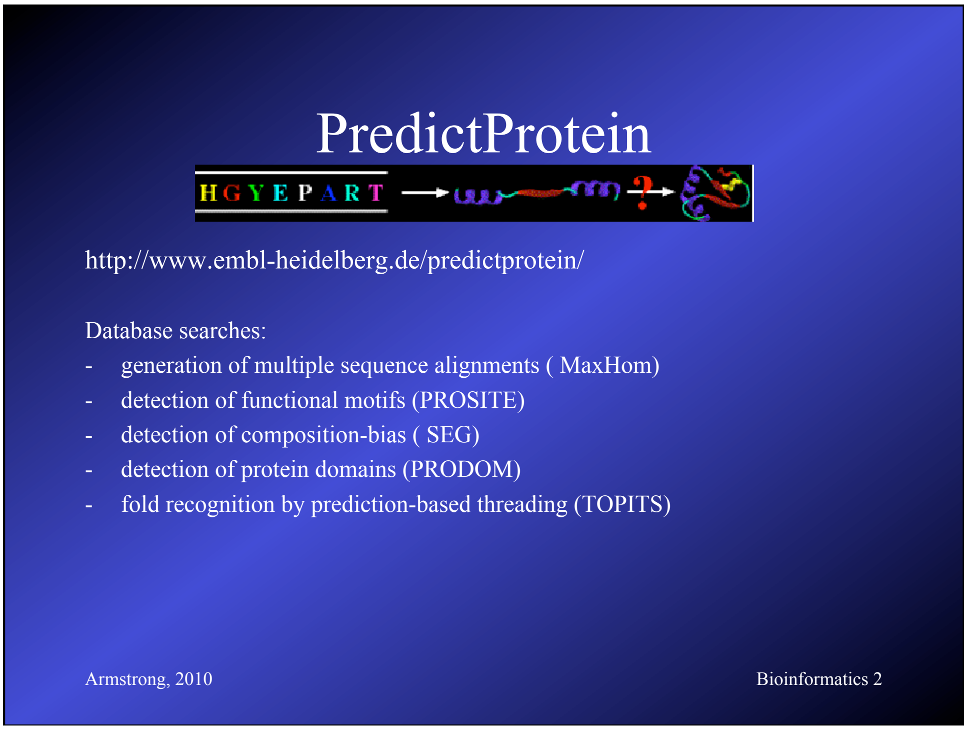


See http://www.ebi.ac.uk/interpro/

InterPro

DATABASE	VERSION	ENTRIES
SWISS-PROT	48	197228
PRINTS	38	1900
TREMBL	31.1	2342938
PFAM	18	7973
PROSITE	19.10	1882

Currently 15 databases, plans to add 3 new ones this month.

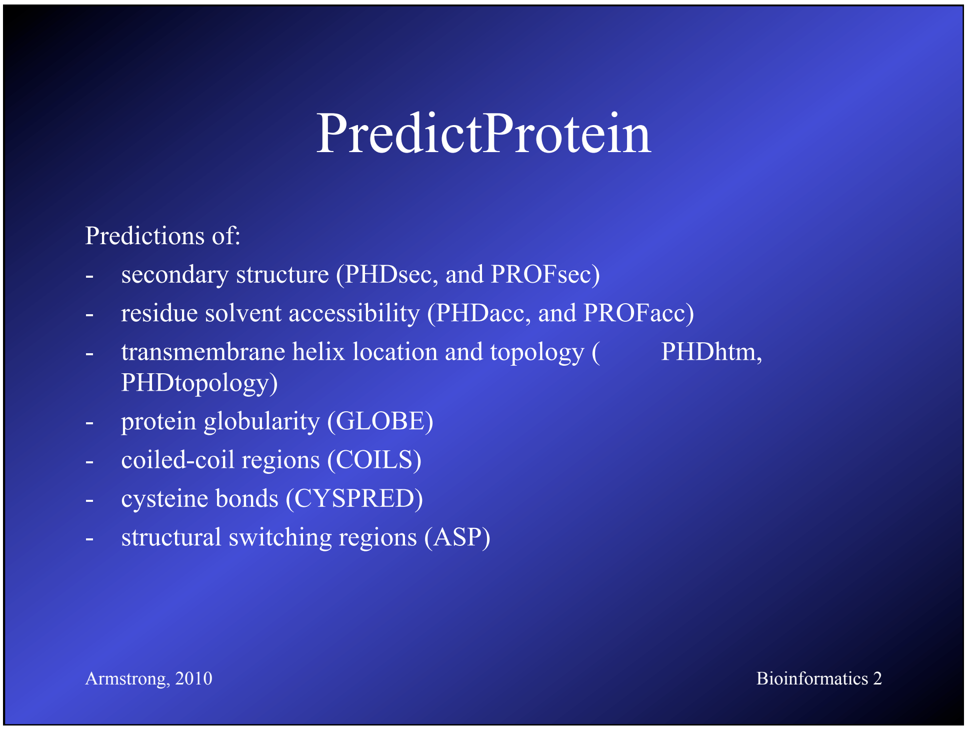


PredictProtein

http://www.embl-heidelberg.de/predictprotein/

Database searches:

- generation of multiple sequence alignments (MaxHom)
- detection of functional motifs (PROSITE)
- detection of composition-bias (SEG)
- detection of protein domains (PRODOM)
- fold recognition by prediction-based threading (TOPITS)

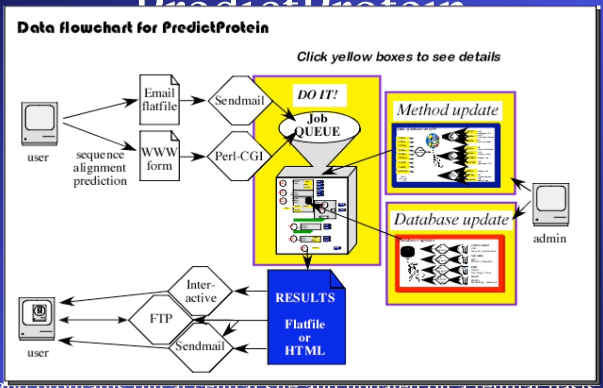


PredictProtein

Predictions of:

- secondary structure (PHDsec, and PROFsec)
- residue solvent accessibility (PHDacc, and PROFacc)
- transmembrane helix location and topology (PHDhtm, PHDtopology)
- protein globularity (GLOBE)
- coiled-coil regions (COILS)
- cysteine bonds (CYSPRED)
- structural switching regions (ASP)

Data and methods in PredictProtein



Add data and programs run at central site and updated on a regular basis

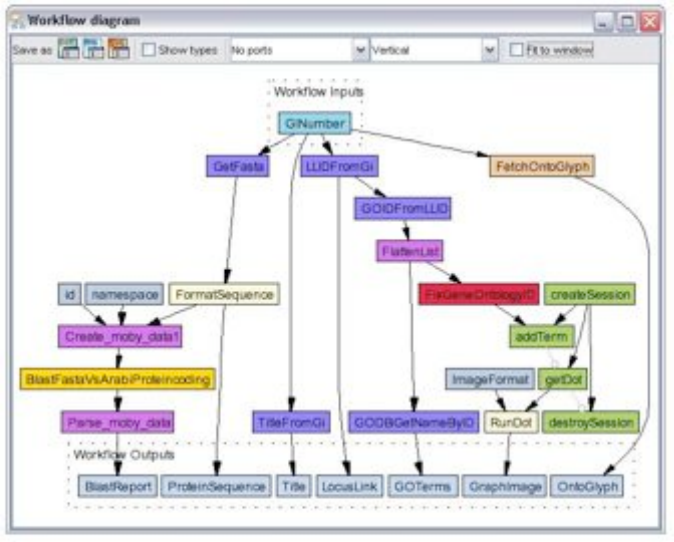
Too many programs/databases

- How do we keep track of our own queries?
  - Repeat an old query
  - Run the same tests on a new sequence
  - Run 100s of sequences..
  - Document the process for a paper or client or for quality assurance

Workflow managers

- Locate and manage connections to software and databases
- Record actions
- Replay a workflow at a later date or against multiple sequences
- Manages redundant external sources (e.g. multiple blast servers)
- Can connect to specialist local sources

- http://taverna.sourceforge.net/
- Open source and free to download
- Runs on PC/linux/mac
- Drag-n-Drop interface to bioinformatics analysis



Workflow diagram

Save as □ Show types No ports Vertical □ Fit to window