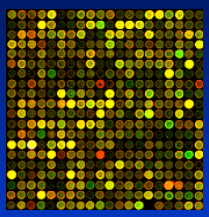
## Protein Bioinformatics



April 24, 2008
Jon Pevsner
pevsner@jhmi.edu
260.841

## Outline for today

BLAST and BLAT

UCSC: Proteome browser, gene sorter, BLAT

NCBI: RefSeq, Trace

UniProt: knowledgebase, UniRef, UniParc

## Dayhoff's 34 protein superfamilies

| Protein | PAMs per 100 million years |
|---|---|
| Ig kappa chain | 37 |
| Kappa casein | 33 |
| luteinizing hormone b | 30 |
| lactalbumin | 27 |
| complement component 3 | 27 |
| epidermal growth factor | 26 |
| proopiomelanocortin | 21 |
| pancreatic ribonuclease | 21 |
| haptoglobin alpha | 20 |
| serum albumin | 19 |
| phospholipase A2, group IB | 19 |
| prolactin | 17 |
| carbonic anhydrase C | 16 |
| Hemoglobin $\alpha$ | 12 |
| Hemoglobin $\beta$ | 12 |

## Dayhoff's 34 protein superfamilies

| Protein | PAMs per 100 million years |
|---|---|
| Ig kappa chain | 37 |
| Kappa casein | 33 |
| luteinizing hormone b | 30 |
| lactalbumin | 27 |
| complement component 3 | 27 |
| epidermal growth factor | 26 |
| proopiomelanocortin | 21 |
| pancreatic ribonuclease | 21 |
| ha | |
| seru | |
| phos | |
| prola | |
| carb | |
| Hem | |
| Hem | |

human (NP_005203) versus mouse (NP_031812)

```
Score = 57.8 bits (138),  Expect = 3e-07
Identities = 39/118 (33%), Positives = 61/118 (51%), Gaps = 2/118 (1%)

Query  1    MKSFLLVVNALALTLPFLAVEVQNQKQPACHENDERPFYQKTAPYVPMYYVPNSYPYYGY  60
            M++F++V+N LALTLPFLA E+QN       E ++  + ++   Y P+  Y N +  Y
Sbjct  2    MRNFIVVMNYLALTLPFLAAEIQNPDSNCRGENDIVYDEQRVLYTPVRSVLN-FNQYEP  60

Query  61   NLYQRPPAI-AINNPYVPRTYYANPAVVRPHAQIPQRQYLPNSHPPYTVVRLPNLHPSF  117
            N Y  BP++ A  +PY+          ++R  A I + Q +PN     V    +PSF
Sbjct  61   NYYHYRPSLPATASPYMYYPLVVRLLLLRSPAPISKWQSMPNFPQSAGVPYAIPNPSF  118
```

## Dayhoff's 34 protein superfamilies

| Protein | PAMs per 100 million years |
|---|---|
| apolipoprotein A-II | 10 |
| lysozyme | 9.8 |
| gastrin | 9.8 |
| myoglobin | 8.9 |
| nerve growth factor | 8.5 |
| myelin basic protein | 7.4 |
| thyroid stimulating hormone b | 7.4 |
| parathyroid hormone | 7.3 |
| parvalbumin | 7.0 |
| trypsin | 5.9 |
| insulin | 4.4 |
| calcitonin | 4.3 |
| arginine vasopressin | 3.6 |
| adenylate kinase 1 | 3.2 |



## Dayhoff's 34 protein superfamilies

| Protein | PAMs per 100 million years |
|---|---|
| triosephosphate isomerase 1 | 2.8 |
| vasoactive intestinal peptide | 2.6 |
| glyceraldehyde phosph. dehydrogease | 2.2 |
| cytochrome c | 2.2 |
| collagen | 1.7 |
| troponin C, skeletal muscle | 1.5 |
| alpha crystallin B chain | 1.5 |
| glucagon | 1.2 |
| glutamate dehydrogenase | 0.9 |
| histone H2B, member Q | 0.9 |
| ubiquitin | 0 |

## Pairwise alignment of human (NP_005203) versus mouse (NP_031812) ubiquitin

```
 Score = 1316 bits (3407),  Expect = 0.0
 Identities = 681/685 (99%), Positives = 682/685 (99%), Gaps = 0/685 (0%)

Query  1    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN  60
            MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN
Sbjct  1    MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYN  60

Query  61   IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI  120
            IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI
Sbjct  61   IQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLI  120

Query  121  FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA  180
            FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA
Sbjct  121  FAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKA  180

Query  181  KIQDKEGIPSDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT  240
            KIQDKEGIP DQQRLIFAGKQLE GRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT
Sbjct  181  KIQDKEGIPPDQQRLIFAGKQLEGGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKT  240

Query  241  ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR  300
            ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR
Sbjct  241  ITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLR  300
```

## Dayhoff's numbers of "accepted point mutations": what amino acid substitutions occur in proteins?

| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly |
|---|---|---|---|---|---|---|---|---|
| A | | | | | | | | |
| R | 30 | | | | | | | |
| N | 109 | 17 | | | | | | |
| D | 154 | 0 | 532 | | | | | |
| C | 33 | 10 | 0 | 0 | | | | |
| Q | 93 | 120 | 50 | 76 | 0 | | | |
| E | 266 | 0 | 94 | 831 | 0 | 422 | | |
| G | 579 | 10 | 156 | 162 | 10 | 30 | 112 | |
| H | 21 | 103 | 226 | 43 | 10 | 243 | 23 | 10 |

Dayhoff (1978) p.346.



## The relative mutability of amino acids

| | | | |
|---|---|---|---|
| **Asn** | **134** | **His** | **66** |
| **Ser** | **120** | **Arg** | **65** |
| **Asp** | **106** | **Lys** | **56** |
| **Glu** | **102** | **Pro** | **56** |
| **Ala** | **100** | **Gly** | **49** |
| **Thr** | **97** | **Tyr** | **41** |
| **Ile** | **96** | **Phe** | **41** |
| **Met** | **94** | **Leu** | **40** |
| **Gln** | **93** | **Cys** | **20** |
| **Val** | **74** | **Trp** | **18** |

Note that alanine is normalized to a value of 100.
Trp and cys are least mutable.
Asn and ser are most mutable.

## Normalized frequencies of amino acids

| | | | |
|---|---|---|---|
| Gly | 8.9% | Arg | 4.1% |
| Ala | 8.7% | Asn | 4.0% |
| Leu | 8.5% | Phe | 4.0% |
| Lys | 8.1% | Gln | 3.8% |
| Ser | 7.0% | Ile | 3.7% |
| Val | 6.5% | His | 3.4% |
| Thr | 5.8% | Cys | 3.3% |
| Pro | 5.1% | Tyr | 3.0% |
| Glu | 5.0% | Met | 1.5% |
| Asp | 4.7% | Trp | 1.0% |

- blue=6 codons; red=1 codon
- These frequencies $f_i$ sum to 1



## Dayhoff's mutation probability matrix for the evolutionary distance of 1 PAM

**We have considered three kinds of information:**
- a table of number of accepted point mutations (PAMs)
- relative mutabilities of the amino acids
- normalized frequencies of the amino acids in PAM data

This information can be combined into a "mutation probability matrix" in which each element $M_{ij}$ gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval (e.g. 1 PAM).



## Dayhoff's PAM1 mutation probability matrix

| | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly | H His |
|---|---|---|---|---|---|---|---|---|---|
| A | 9867 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 |
| R | 1 | 9913 | 1 | 0 | 1 | 10 | 0 | 0 | 10 |
| N | 4 | 1 | 9822 | 36 | 0 | 4 | 6 | 6 | 21 |
| D | 6 | 0 | 42 | 9859 | 0 | 6 | 53 | 6 | 4 |
| C | 1 | 1 | 0 | 0 | 9973 | 0 | 0 | 0 | 1 |
| Q | 3 | 9 | 4 | 5 | 0 | 9876 | 27 | 1 | 23 |
| E | 10 | 0 | 7 | 56 | 0 | 35 | 9865 | 4 | 2 |
| G | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 9935 | 1 |
| H | 1 | 8 | 18 | 3 | 1 | 20 | 1 | 0 | 9912 |
| I | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 |

Each element of the matrix shows the probability that an original amino acid (top) will be replaced by another amino acid (side)

## Dayhoff's PAM0 mutation probability matrix: the rules for extremely slowly evolving proteins

| PAM0 | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu |
|---|---|---|---|---|---|---|---|
| A | 100% | 0% | 0% | 0% | 0% | 0% | 0% |
| R | 0% | 100% | 0% | 0% | 0% | 0% | 0% |
| N | 0% | 0% | 100% | 0% | 0% | 0% | 0% |
| D | 0% | 0% | 0% | 100% | 0% | 0% | 0% |
| C | 0% | 0% | 0% | 0% | 100% | 0% | 0% |
| Q | 0% | 0% | 0% | 0% | 0% | 100% | 0% |
| E | 0% | 0% | 0% | 0% | 0% | 0% | 100% |
| G | 0% | 0% | 0% | 0% | 0% | 0% | 0% |

Top: original amino acid
Side: replacement amino acid



## Dayhoff's PAM2000 mutation probability matrix: the rules for very distantly related proteins

| PAM∞ | A Ala | R Arg | N Asn | D Asp | C Cys | Q Gln | E Glu | G Gly |
|---|---|---|---|---|---|---|---|---|
| A | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7% | 8.7 |
| R | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1% | 4.1 |
| N | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0% | 4.0 |
| D | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7% | 4.7 |
| C | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3% | 3.3 |
| Q | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8% | 3.8 |
| E | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0% | 5.0 |
| G | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9% | 8.9 |

Top: original amino acid
Side: replacement amino acid



## PAM250 mutation probability matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 13 | 6 | 9 | 9 | 5 | 8 | 9 | 12 | 6 | 8 | 6 | 7 | 7 | 4 | 11 | 11 | 11 | 2 | 4 | 9 |
| R | 3 | 17 | 4 | 3 | 2 | 5 | 3 | 2 | 6 | 3 | 2 | 9 | 4 | 1 | 4 | 4 | 3 | 7 | 2 | 2 |
| N | 4 | 4 | 6 | 7 | 2 | 5 | 6 | 4 | 6 | 3 | 2 | 5 | 3 | 2 | 4 | 5 | 4 | 2 | 3 | 3 |
| D | 5 | 4 | 8 | 11 | 1 | 7 | 10 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| C | 2 | 1 | 1 | 1 | 52 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 4 | 2 |
| Q | 3 | 5 | 5 | 6 | 1 | 10 | 7 | 3 | 7 | 2 | 3 | 5 | 3 | 1 | 4 | 3 | 3 | 1 | 2 | 3 |
| E | 5 | 4 | 7 | 11 | 1 | 9 | 12 | 5 | 6 | 3 | 2 | 5 | 3 | 1 | 4 | 5 | 5 | 1 | 2 | 3 |
| G | 12 | 5 | 10 | 10 | 4 | 7 | 9 | 27 | 5 | 5 | 4 | 6 | 5 | 3 | 8 | 11 | 9 | 2 | 3 | 7 |
| H | 2 | 5 | 5 | 4 | 2 | 7 | 4 | 2 | 15 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 |
| I | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 10 | 6 | 2 | 6 | 5 | 2 | 3 | 4 | 1 | 3 | 9 |
| L | 6 | 4 | 4 | 3 | 2 | 6 | 4 | 3 | 5 | 15 | 34 | 4 | 20 | 13 | 5 | 4 | 6 | 6 | 7 | 13 |
| K | 6 | 18 | 10 | 8 | 2 | 10 | 8 | 5 | 8 | 5 | 4 | 24 | 9 | 2 | 6 | 8 | 8 | 4 | 3 | 5 |
| M | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 2 | 3 | 2 | 6 | 2 | 1 | 1 | 1 | 1 | 1 | 2 |
| F | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 5 | 6 | 1 | 4 | 32 | 1 | 2 | 2 | 4 | 20 | 3 |
| P | 7 | 5 | 5 | 4 | 3 | 5 | 4 | 5 | 5 | 3 | 3 | 4 | 3 | 2 | 20 | 6 | 5 | 1 | 2 | 4 |
| S | 9 | 6 | 8 | 7 | 7 | 6 | 7 | 9 | 6 | 5 | 4 | 7 | 5 | 3 | 9 | 10 | 9 | 4 | 4 | 6 |
| T | 8 | 5 | 6 | 6 | 4 | 5 | 5 | 6 | 4 | 6 | 4 | 6 | 5 | 3 | 6 | 8 | 11 | 2 | 3 | 6 |
| W | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 55 | 1 | 0 |
| Y | 1 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 15 | 1 | 2 | 2 | 3 | 31 | 2 |
| V | 7 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 4 | 15 | 10 | 4 | 10 | 5 | 5 | 5 | 7 | 2 | 4 | 17 |

Top: original amino acid
Side: replacement amino acid

## PAM250 log odds scoring matrix

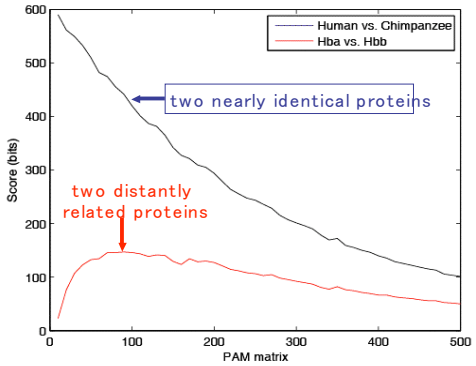
| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 2 | | | | | | | | | | | | | | | | | | | |
| **R** | -2 | 6 | | | | | | | | | | | | | | | | | | |
| **N** | 0 | 0 | 2 | | | | | | | | | | | | | | | | | |
| **D** | 0 | -1 | 2 | 4 | | | | | | | | | | | | | | | | |
| **C** | -2 | -4 | -4 | -5 | 12 | | | | | | | | | | | | | | | |
| **Q** | 0 | 1 | 1 | 2 | -5 | 4 | | | | | | | | | | | | | | |
| **E** | 0 | -1 | 1 | 3 | -5 | 2 | 4 | | | | | | | | | | | | | |
| **G** | 1 | -3 | 0 | 1 | -3 | -1 | 0 | 5 | | | | | | | | | | | | |
| **H** | -1 | 2 | 2 | 1 | -3 | 3 | 1 | -2 | 6 | | | | | | | | | | | |
| **I** | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -2 | 5 | | | | | | | | | | |
| **L** | -2 | -3 | -3 | -4 | -6 | -2 | -3 | -4 | -2 | -2 | 6 | | | | | | | | | |
| **K** | -1 | 3 | 1 | 0 | -5 | 1 | 0 | -2 | 0 | -2 | -3 | 5 | | | | | | | | |
| **M** | -1 | 0 | -2 | -3 | -5 | -1 | -2 | -3 | -2 | 2 | 4 | 0 | 6 | | | | | | | |
| **F** | -3 | -4 | -3 | -6 | -4 | -5 | -5 | -5 | -2 | 1 | 2 | -5 | 0 | 9 | | | | | | |
| **P** | 1 | 0 | 0 | -1 | -3 | 0 | -1 | 0 | 0 | -2 | -3 | -1 | -2 | -5 | 6 | | | | | |
| **S** | 1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | -1 | -1 | -3 | 0 | -2 | -3 | 1 | 2 | | | | |
| **T** | 1 | -1 | 0 | 0 | -2 | -1 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | -3 | 0 | 1 | 3 | | | |
| **W** | -6 | 2 | -4 | -7 | -8 | -5 | -7 | -7 | -3 | -5 | -2 | -3 | -4 | 0 | -6 | -2 | -5 | 17 | | |
| **Y** | -3 | -4 | -2 | -4 | 0 | -4 | -4 | -5 | 0 | -1 | -1 | -4 | -2 | 7 | -5 | -3 | -3 | 0 | 10 | |
| **V** | 0 | -2 | -2 | -2 | -2 | -2 | -2 | -1 | -2 | 4 | 2 | -2 | 2 | -1 | -1 | -1 | 0 | -6 | -2 | 4 |



## PAM10 log odds scoring matrix

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 7 | | | | | | | | | | | | | | | | | | | |
| **R** | -10 | 9 | | | | | | | | | | | | | | | | | | |
| **N** | -7 | -9 | 9 | | | | | | | | | | | | | | | | | |
| **D** | -6 | -17 | -1 | 8 | | | | | | | | | | | | | | | | |
| **C** | -10 | -11 | -17 | -21 | 10 | | | | | | | | | | | | | | | |
| **Q** | -7 | -4 | -7 | -6 | -20 | 9 | | | | | | | | | | | | | | |
| **E** | -5 | -15 | -5 | 0 | -20 | -1 | 8 | | | | | | | | | | | | | |
| **G** | -4 | -13 | -6 | -6 | -13 | -10 | -7 | 7 | | | | | | | | | | | | |
| **H** | -11 | -4 | -2 | -7 | -10 | -2 | -9 | -13 | 10 | | | | | | | | | | | |
| **I** | -8 | -8 | -8 | -11 | -9 | -11 | -8 | -17 | -13 | 9 | | | | | | | | | | |
| **L** | -9 | -12 | -10 | -19 | -21 | -8 | -13 | -14 | -9 | -4 | 7 | | | | | | | | | |
| **K** | -10 | -2 | -4 | -8 | -20 | -6 | -7 | -10 | -10 | -9 | -11 | 7 | | | | | | | | |
| **M** | -8 | -7 | -15 | -17 | -20 | -7 | -10 | -12 | -17 | -3 | -2 | -4 | 12 | | | | | | | |
| **F** | -12 | -12 | -12 | -21 | -19 | -19 | -20 | -12 | -9 | -5 | -5 | -20 | -7 | 9 | | | | | | |
| **P** | -4 | -7 | -9 | -12 | -11 | -6 | -9 | -10 | -7 | -12 | -10 | -10 | -11 | -13 | 8 | | | | | |
| **S** | -3 | -6 | -2 | -7 | -6 | -8 | -7 | -4 | -9 | -10 | -12 | -7 | -8 | -9 | -4 | 7 | | | | |
| **T** | -3 | -10 | -5 | -8 | -11 | -9 | -9 | -10 | -11 | -5 | -10 | -6 | -7 | -12 | -7 | -2 | 6 | | | |
| **W** | -20 | -5 | -11 | -21 | -22 | -19 | -23 | -21 | -10 | -20 | -9 | -18 | -19 | -7 | -20 | -8 | -10 | 13 | | |
| **Y** | -11 | -14 | -7 | -17 | -7 | -18 | -11 | -20 | -6 | -9 | -10 | -12 | -17 | -1 | -20 | -10 | -5 | -8 | 10 | |
| **V** | -5 | -11 | -12 | -11 | -9 | -10 | -10 | -9 | -9 | -1 | -5 | -13 | -4 | -12 | -9 | -10 | -6 | -22 | -10 | 8 |





## BLAST

BLAST (Basic Local Alignment Search Tool) allows rapid sequence comparison of a query sequence against a database.

The BLAST algorithm is fast, accurate, and web-accessible.



## Four components to a BLAST search

(1) Choose the sequence (query)

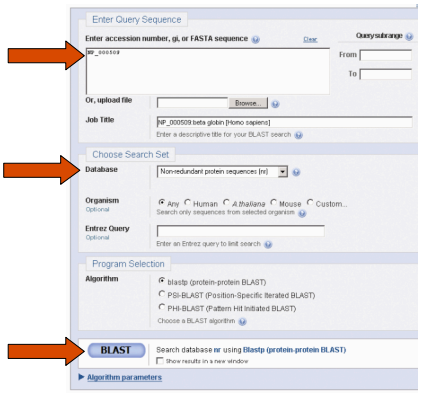
(2) Select the BLAST program

(3) Choose the database to search

(4) Choose optional parameters

Then click “BLAST”





## Choose the BLAST program

| Program | Input | | Database |
|---|---|---|---|
| blastn | DNA | 1 | DNA |
| blastp | protein | 1 | protein |
| blastx | DNA | 6 | protein |
| tblastn | protein | 6 | DNA |
| tblastx | DNA | 36 | DNA |

Fig. 4.3
page 91

## How a BLAST search works

"The central idea of the BLAST algorithm is to confine attention to segment pairs that contain a word pair of length w with a score of at least T."

Altschul et al. (1990)

(page 101, 102)

## How the original BLAST algorithm works: three phases

Phase 1: compile a list of word pairs (w=3) above threshold T

Example: for a human RBP query …FSGTWYA… (query word is in yellow)

A list of words (w=3) is:

```
FSG SGT GTW TWY WYA
YSG TGT ATW SWY WFA
FTG SVT GSW TWF WYS
```

Fig. 4.13
page 101

## Phase 1: compile a list of words (w=3)

neighborhood word hits > threshold

(T=11)

neighborhood word hits < below threshold

| | | |
|---|---|---|
| GTW | 6,5,11 | 22 |
| GSW | 6,1,11 | 18 |
| ATW | 0,5,11 | 16 |
| NTW | 0,5,11 | 16 |
| GTY | 6,5,2 | 13 |
| GNW | | 10 |
| GAW | | 9 |

Fig. 4.13
page 101

## Pairwise alignment scores are determined using a scoring matrix such as Blosum62

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

Page 61

## How a BLAST search works: 3 phases

Phase 2:

Scan the database for entries that match the compiled list.

This is fast and relatively easy.

Fig. 4.13
page 101

## How a BLAST search works: 3 phases

Phase 3: when you manage to find a hit (i.e. a match between a "word" and a database entry), extend the hit in either direction.

Keep track of the score (use a scoring matrix)

Stop when the score drops below some cutoff.

```
KENFDKARFSGTWYAMAKKDPEG 50  RBP (query)
MKGLDIQKVAGTWYSLAMAASD. 44  lactoglobulin (hit)
```

extend Hit! extend



## How a BLAST search works: threshold

You can modify the threshold parameter.

The default value for blastp is 11.

To change it, enter "-f 16" or "-f 5" in the advanced options of NetBLAST.

(To find NetBLAST enter it as a query on the NCBI site map.)



## Phase 1: compile a list of words (w=3)

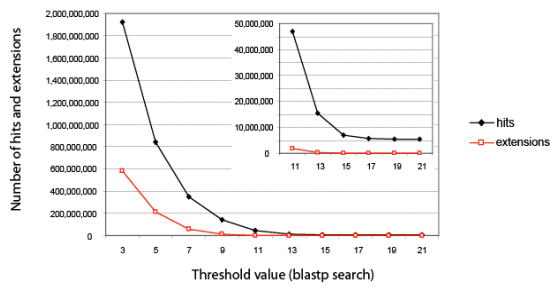
neighborhood word hits > threshold

(T=11)

neighborhood word hits < below threshold

| | | |
|---|---|---|
| GTW | 6,5,11 | 22 |
| ASW | 6,1,11 | 18 |
| ATW | 0,5,11 | 16 |
| NTW | 0,5,11 | 16 |
| GTY | 6,5,2 | 13 |
| GNW | | 10 |
| GAW | | 9 |

Fig. 4.13
page 101




2e Fig. 4.12

## How to interpret a BLAST search: expect value

The expect value *E* is the number of alignments with scores greater than or equal to score *S* that are expected to occur by chance in a database search.

An *E* value is related to a probability value *p*.

The key equation describing an *E* value is:

$$E = Kmn\, e^{-\lambda S}$$

page 105

## $E = Kmn\, e^{-\lambda S}$

This equation is derived from a description of the extreme value distribution

*S* = the score

*E* = the expect value = the number of high-scoring segment pairs (HSPs) expected to occur with a score of at least *S*

*m*, *n* = the length of two sequences

λ, *K* = Karlin Altschul statistics

page 105

## How to interpret BLAST: *E* values and *p* values

Very small *E* values are very similar to *p* values.
*E* values of about 1 to 10 are far easier to interpret than corresponding *p* values.

| *E* | *p* |
|---|---|
| 10 | 0.99995460 |
| 5 | 0.99326205 |
| 2 | 0.86466472 |
| 1 | 0.63212056 |
| 0.1 | 0.09516258 (about 0.1) |
| 0.05 | 0.04877058 (about 0.05) |
| 0.001 | 0.00099950 (about 0.001) |
| 0.0001 | 0.0001000 |

Table 4.4
page 107

## Outline for today

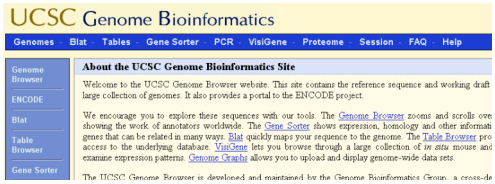
BLAST and BLAT

UCSC: Proteome browser, gene sorter, BLAT

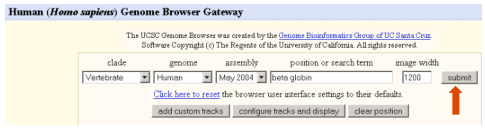
NCBI: RefSeq, Trace

UniProt: knowledgebase, UniRef, UniParc

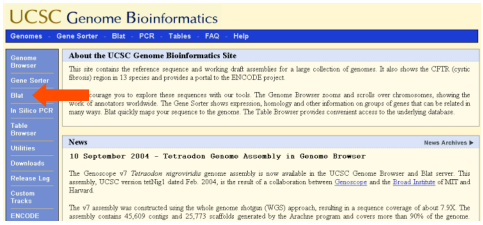
[1] Visit http://genome.ucsc.edu/, click Genome Browser



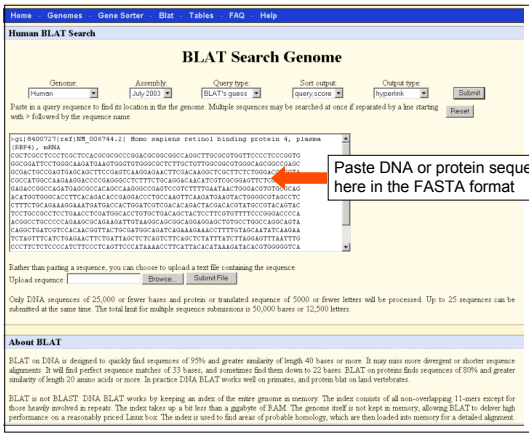

[2] Choose organisms, enter query (beta globin), hit submit



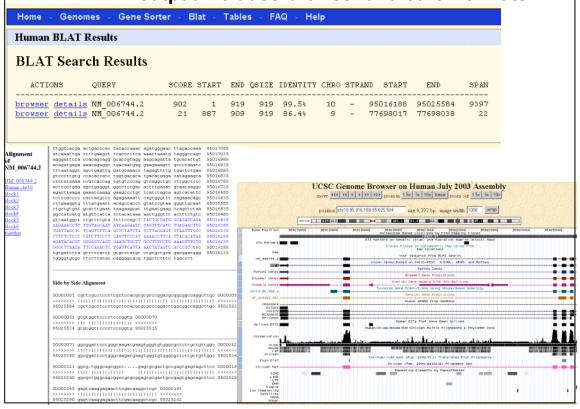
## To access BLAT, visit http://genome.ucsc.edu




"BLAT on DNA is designed to quickly find sequences of 95% and greater similarity of length 40 bases or more. It may miss more divergent or shorter sequence alignments. It will find perfect sequence matches of 33 bases, and sometimes find them down to 20 bases. BLAT on proteins finds sequences of 80% and greater similarity of length 20 amino acids or more. In practice DNA BLAT works well on primates, and protein blat on land vertebrates." --BLAT website




## BLAT output includes browser and other formats



Blastz

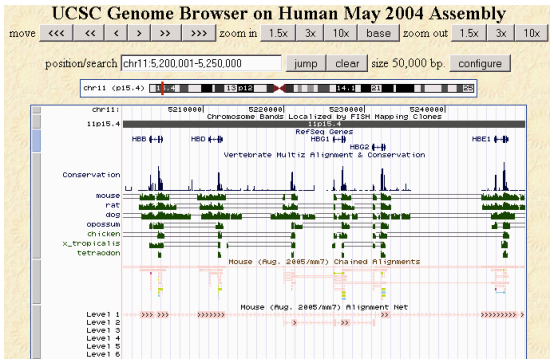



Fig. 5.17

## Outline for today

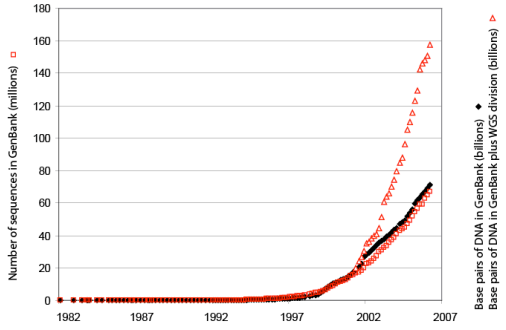
BLAST and BLAT

UCSC: Proteome browser, gene sorter, BLAT

NCBI: RefSeq, Trace

UniProt: knowledgebase, UniRef, UniParc

## Growth of GenBank + Whole Genome Shotgun



## Accession numbers are labels for sequences

NCBI includes databases (such as GenBank) that contain information on DNA, RNA, or protein sequences.
You may want to acquire information beginning with a query such as the name of a protein of interest, or the raw nucleotides comprising a DNA sequence of interest.

DNA sequences and other molecular data are tagged with accession numbers that are used to identify a sequence or other record relevant to molecular data.



## What is an accession number?

An accession number is label that used to identify a sequence. It is a string of letters and/or numbers that corresponds to a molecular sequence.

Examples (all for retinol-binding protein, RBP4):

| | | |
|---|---|---|
| X02775 | GenBank genomic DNA sequence | **DNA** |
| NT_030059 | Genomic contig | |
| Rs7079946 | dbSNP (single nucleotide polymorphism) | |
| N91759.1 | An expressed sequence tag (1 of 170) | **RNA** |
| NM_006744 | RefSeq DNA sequence (from a transcript) | |
| NP_007635 | RefSeq protein | **protein** |
| AAC02945 | GenBank protein | |
| Q28369 | SwissProt protein | |
| 1KT7 | Protein Data Bank structure record | |



## NCBI's important RefSeq project: best representative sequences

RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.

RefSeq identifiers include the following formats:

| | |
|---|---|
| Complete genome | NC_###### |
| Complete chromosome | NC_###### |
| Genomic contig | NT_###### |
| mRNA (DNA format) | NM_###### e.g. NM_006744 |
| Protein | NP_###### e.g. NP_006735 |

## NCBI's RefSeq project: accession for genomic, mRNA, protein sequences

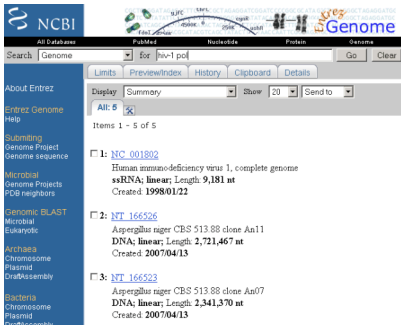
| Accession | Molecule | Method | Note |
|---|---|---|---|
| AC_123456 | Genomic | Mixed | Alternate complete genomic |
| AP_123456 | Protein | Mixed | Protein products; alternate |
| NC_123456 | Genomic | Mixed | Complete genomic molecules |
| NG_123456 | Genomic | Mixed | Incomplete genomic regions |
| NM_123456 | mRNA | Mixed | Transcript products; mRNA |
| NM_123456789 | mRNA | Mixed | Transcript products; 9-digit |
| NP_123456 | Protein | Mixed | Protein products; |
| NP_123456789 | Protein | Curation | Protein products; 9-digit |
| NR_123456 | RNA | Mixed | Non-coding transcripts |
| NT_123456 | Genomic | Automated | Genomic assemblies |
| NW_123456 | Genomic | Automated | Genomic assemblies |
| NZ_ABCD12345678 | Genomic | Automated | Whole genome shotgun data |
| XM_123456 | mRNA | Automated | Transcript products |
| XP_123456 | Protein | Automated | Protein products |
| XR_123456 | RNA | Automated | Transcript products |
| YP_123456 | Protein | Auto. & Curated | Protein products |
| ZP_12345678 | Protein | Automated | Protein products |

## Example of how to access sequence data: HIV-1 *pol*

There are many possible approaches. Begin at the main page of NCBI, and type an Entrez query: hiv-1 pol



## Searching for HIV-1 *pol*: Following the "genome" link yields a manageable five results



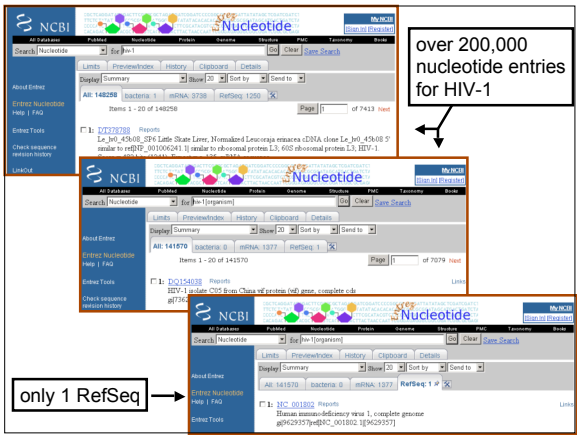



## Example of how to access sequence data: HIV-1 *pol*

For the Entrez query: hiv-1 pol
there are about 80,000 nucleotide or protein records
(and >200,000 records for a search for "hiv-1"),
but these can easily be reduced in two easy steps:

--specify the organism, e.g. hiv-1[organism]
--limit the output to RefSeq!



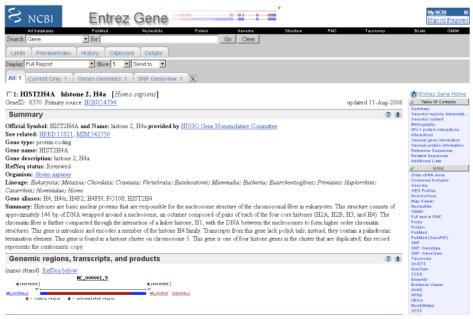


## Examples of how to access sequence data: histone

| query for “histone” | # results |
|---|---|
| protein records | 21847 |
| RefSeq entries | 7544 |
| RefSeq (limit to human) | 1108 |
| NOT deacetylase | 697 |

At this point, select a reasonable candidate (e.g. histone 2, H4) and follow its link to Entrez Gene. There, you can confirm you have the right gene/protein.

8-12-06




## Outline for today

BLAST and BLAT

UCSC: Proteome browser, gene sorter, BLAT

NCBI: RefSeq, Trace

UniProt: knowledgebase, UniRef, UniParc


UniProt
the universal protein resource
Text Search UniProt Knowledgebase
Home
About UniProt
Getting Started
Searches/Tools
Databases
Support/Documentation
Notice: This site will be replaced with beta.uniprot.org. Please send u
Text Search
BLAST
FAQ
Help Desk
Download
Welcome to UniProt
UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.
UniProt has three components, each optimized for different uses. The UniProt Knowledgebase (UniProtKB) is the central access point for extensive curated protein information, including function, classification, and cross-reference. The UniProt Reference Clusters (UniRef) databases combine closely related sequences into a single record to speed searches. The UniProt Archive (UniParc) is a comprehensive repository, reflecting the history of all protein sequences.
The sequences and information in UniProt are accessible via text search, BLAST similarity search, and FTP.
European Bioinformatics Institute
Swiss Institute of Bioinformatics
Georgetown University


Search in
Query
Protein Knowledgebase (UniProtKB)
Search
Clear
Fields »
Search
Blast
WELCOME
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
What we provide
UniProtKB
Protein knowledgebase, consists of two sections:
Swiss-Prot, which is manually annotated and reviewed.
TrEMBL, which is automatically annotated and is not reviewed.
UniRef
Sequence clusters, used to speed up similarity searches.
UniParc
Sequence archive, used to keep track of sequences and their identifiers.
Supporting data
Literature citations, taxonomy, keywords and more.
UniProt