CAP 5510-6
BLAST

BIOINFORMATICS

Su-Shing Chen

CISE

BLAST

Basic Local Alignment Search Tool

Prof Su-Shing Chen

A Fast Pair-wise Alignment and Database Searching Tool

What is BLAST ?

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
- BLAST provides a method for rapid searching of nucleotide and protein databases.

Why is BLAST used ?

- Sequence alignments provide a powerful way to compare novel sequences with previously characterized genes
- Infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

How does BLAST work ?

- Finds regions of local as well as global similarity between sequences.
- Compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches.
- The Similarity may provide important clues to the function of uncharacterized proteins or nucleotides.

Advantages of using BLAST

- The BLAST algorithm was written balancing speed and increased sensitivity for distant sequence relationships.
- BLAST emphasizes regions of local alignment to detect relationships among sequences which share only isolated regions of similarity.

BLAST programs

- The BLAST search allows us to select from several different programs.

BLAST programs contd..

Program	Description
blastp	Compares an amino acid query sequence against a protein sequence database.
blastn	Compares a nucleotide query sequence against a nucleotide sequence database.

BLAST programs contd..

Program	Description
blastx	Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.

BLAST programs contd..

Program	Description
tblastn	Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.

BLAST programs contd..

Program	Description
tblastx	Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Please note that the tblastx program cannot be used with the nr database on the BLAST Web page because it is computationally intensive.

Using BLAST

- Go to :- http://www.ncbi.nlm.nih.gov/BLAST/
- Select the blast program
- Enter the protein/nucleotide in proper search format
- Select Database
- Click on BLAST

Search Format

- The BLAST 'Search' box accepts a number of different types of input and automatically determines the format.
- FASTA format
- Bare Sequence format
- Identifiers
- Go to :- http://www.ncbi.nlm.nih.gov/blast/html/search.html

Search Format contd..

- FASTA format
- A sequence in FASTA format begins with a single-line description, followed by lines of sequence data.
- >gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)

  ```
  QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAED
  TREMPFHVTKQESKPVQMMCMNNSFNVATLPAEKM
  KILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTE
  WTNPNTMEKRRVKVYLPQMKIEEKYNLTSVLMALG
  MTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGI
  EMAGSTGVIEDIKHSPESEQFRADHPFLFLIKHNPTN
  TIVYFGRYWSP
  ```

Search Format contd..

- Bare Sequence
- This may be just lines of sequence data, without the FASTA definition line
- It can also be sequence interspersed with numbers and/or spaces, such as the sequence portion of a GenBank/GenPept flatfile report:
- 1 qikdllvsss tdldttlvlv naiyfkgmwk tafnaedtre mpfhvtkqes kpvqmmcmnn
- 61 sfnvatlpae kmkilelpfa sgdlsmlvll pdevsdleri ektinfeklt ewtnpntmek
- 121 rrvkvylpqm kieekynlts vlmalgmtdl fipsanltgi ssaeslkisq avhgafmels
- 181 edgiemagst gviedikhsp eseqfradhp flflikhnpt ntivyfgryw sp

Search Format contd..

- Identifiers
- Normally these are simply accession, accession.version or gi's (e.g., p01013, AAA68881.1, 129295)

BLAST score

- The BLAST score is scoring a local alignment without gaps which consists simply of a pair of equal length segments, one from each of the two sequences being compared.
- The database sequences are assumed to be evolutionary unrelated, i.e. independent of one another.
- A high BLAST score indicates that two proteins are extremely similar in primary structure, much more similar than would be expected by chance.

Statistical Analysis

- To assess whether a given alignment constitutes evidence for homology, it helps to know how strong an alignment can be expected from chance alone.
- “Chance" can mean the comparison of :
  - real but non-homologous sequences
  - real sequences that are shuffled to preserve compositional properties
  - sequences that are generated randomly based upon a DNA or protein sequence model.

Statistics of local sequence comparison

- Statistics for the scores of local alignments are well understood, particularly for local alignments lacking gaps
- HSPs : high-scoring segment pairs. All segment pairs whose scores can not be improved by extension or trimming.
- In the limit of sufficiently large sequence lengths *m* and *n*, the statistics of HSP scores are characterized by two parameters, *K* and *lambda*.

Statistics of local sequence comparison

- E-value : The expected number of HSPs with score at least *S* is given by :-

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

- This is the expected number of times a given score will be exceeded in a databank search
- The parameters *K* and *lambda* can be thought of simply as natural scales for the search space size and the scoring system respectively.

Statistics of local sequence comparison

- In BLAST, FASTA score thresholds are set so that the expected number of reported false positives is fixed (e.g. at E-value=1)
- As the
  - size of the databanks increases,
  - the E-value score threshold increases
  - the reported E-value of a particular comparison will increase
    - And its apparent significance will decrease

Statistics of global sequence comparison

- Very little is known about the random distribution of optimal global alignment scores.
- One of the few methods available for assessing the statistical significance of a particular global alignment is to generate many random sequence pairs of the appropriate length and composition, and calculate the optimal alignment score for each

Bit scores

- Raw scores have little meaning without detailed knowledge of the scoring system used, or more simply its statistical parameters *K* and *lambda*.
- By normalizing a raw score one attains a "bit score" *S'*

$$S' = \frac{\lambda S - \log(K)}{\log(2)}$$

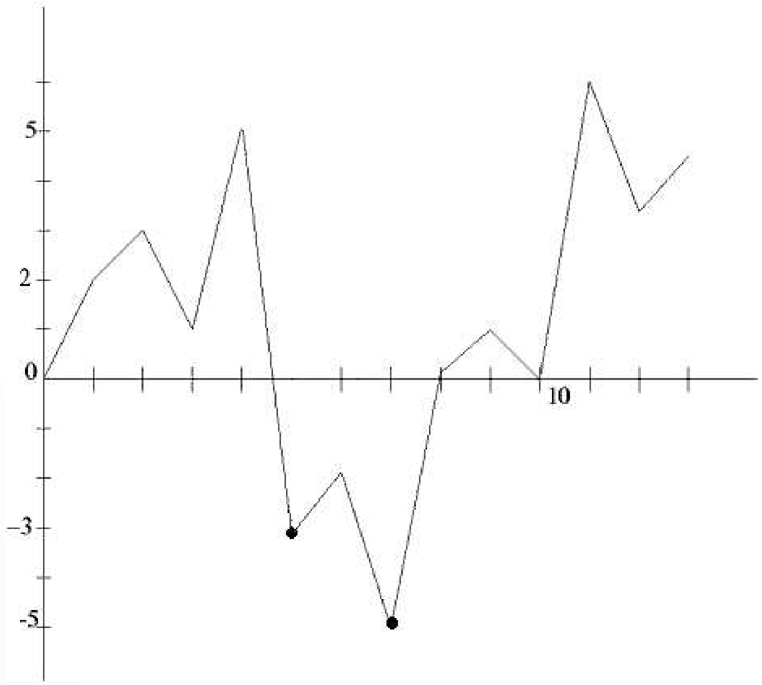
- E-value of bit score

$$E = mn2^{-S'}$$

P-values

- P-value : The probability of finding at least one such HSP with score ≥ S
- P-value = Prob($Y_{max} \geq y$)
- P-value = 1 – e^{-E}

Random Walks



- Ladder Points and Excursions
- BLAST theory focuses on maximum heights achieved by these excursions.

Random Walks

- In case of BLAST :
  - h = 0
- a = -1
  - b = ?
  - Mean step size , E(S) is negative
- Walk is destined eventually to reach -1.

Database searches

- The *E*-value applies to the comparison of two proteins of lengths *m* and *n*.
- How does one assess the significance of an alignment that arises from the comparison of a protein of length *m* to a database containing many different proteins, of varying lengths?
- There are 2 views.

Database searches

- A query is *a priori* more likely to be related to a long than to a short sequence, because long sequences are often composed of multiple distinct domains.
- The BLAST programs take this approach to calculating database *E*-value.
- Assume the *a priori* chance of relatedness is proportional to sequence length then the pairwise *E*-value involving a database sequence of length *n* should be multiplied by *N/n*
  - Where *N* is the total length of the database in residues.

Database searches

- Another view is that all proteins in the database are *a priori* equally likely to be related to the query.
- This implies that a low *E*-value for an alignment involving a short database sequence should carry the same weight as a low *E*-value for an alignment involving a long database sequence.
- Recent versions of the FASTA protein comparison programs take this approach.
- To calculate a "database search" *E*-value, one simply multiplies the pairwise-comparison *E*-value by the number of sequences in the database.

The statistics of gapped alignments

- For ungapped alignments, the statistical parameters can be calculated, using analytic formulas, from the substitution scores and the background residue frequencies of the sequences being compared.
- For gapped alignments, these parameters must be estimated from a large-scale comparison of "random" sequences.
- Some database search programs, such as FASTA or various implementation of the Smith-Waterman algorithm , produce optimal local alignment scores for the comparison of the query sequence to every sequence in the database.
- Most of these scores involve unrelated sequences, and therefore can be used to estimate *lambda* and *K*

The choice of substitution scores

- The scores of any substitution matrix with negative expected score can be written uniquely in the form

$$S_{ij} = \left(\ln \frac{q_{ij}}{p_i \; p_j}\right) / \lambda$$

S_{ij} is a log likelihood ratio.

- Where
  - ⋆ *q_{ij} are* called target frequencies
  - ⋆ *p_i* are background frequencies for the various residues
  - ⋆ *lambda* is a positive constant

The choice of substitution scores

- Multiplying all the scores in a substitution matrix by a positive constant does not change their essence: an alignment that was optimal using the original scores remains optimal.
- Such multiplication alters the parameter *lambda* but not the target frequencies q_{ij}.
- Thus, up to a constant scaling factor, every substitution matrix is uniquely determined by its target frequencies.

Edge Effects

- The statistics described above tend to be somewhat conservative for short sequences.
- The theory supporting these statistics is an asymptotic one, which assumes an optimal local alignment can begin with any aligned pair of residues.
- However, a high-scoring alignment must have some length, and therefore can not begin near to the end of either of two sequences being compared .
- This "edge effect" may be corrected for by calculating an "effective length" for sequences ; the BLAST programs implement such a correction.
- For sequences longer than about 200 residues the edge effect correction is usually negligible.

Edge Effects

- Effective Lengths :-

$$N' = N - (\lambda Y_{max})/H$$

- N is the original length
- Lambda is calculated from :

$$\sum_{i,j} p_i p_j e^{-\lambda s_{ij}} = 1$$

- Relative Entropy H :

$$\Sigma_{i,j}\, q_{ij} log(q_{ij} / (p_i p_j)$$

BLAST printouts

- Some values given in BLAST output :-
  - Score(Y_{max}), Expect, P-value
  - Subs matrix, λ, K, H
    - $K = (C/A)e^{-\lambda}$
    - C is a constant which depends on the Substitution matrix being used and the amino acid frequency
    - A is the mean number of steps taken until the random walk reaches -1

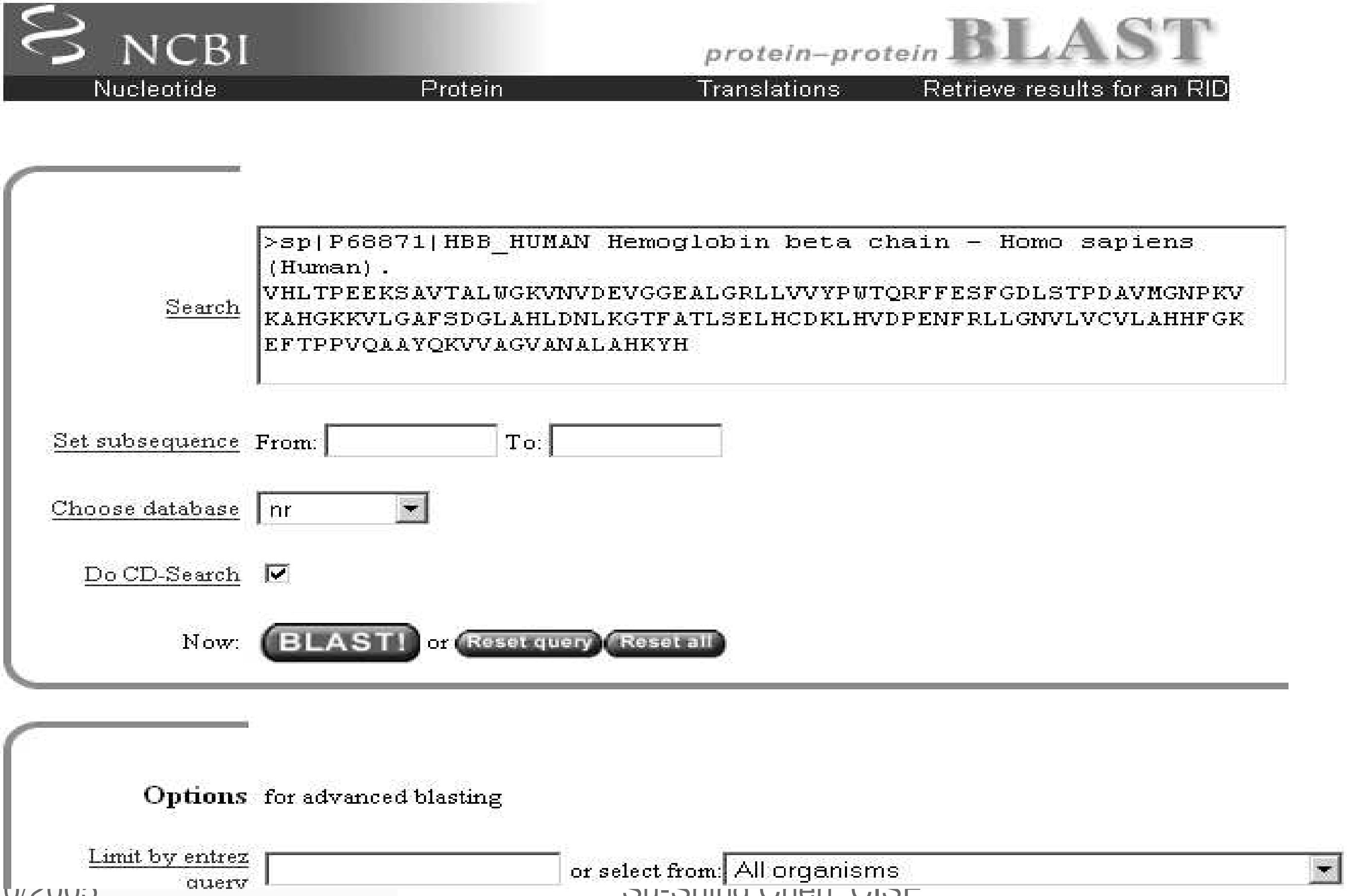
BLAST printouts

- Approximate verification of values displayed in the blast output :
  - Calculate N_1’ and N_2’ from N_1 and N_2
  - Calculate E’ = N_1’ N_2’$Ke^{-\lambda ymax}$
  - Calculate E = 2E’
  - Calculate Expect = $(1-e^{-E})D/N_2$
  - Calculate P-value = $1-e^{-Expect}$

BLAST example - Input

- >sp|P68871|HBB_HUMAN Hemoglobin
  VHLTPEEKSAVTALWGKVNVDEVGGEALGRLL
  VVYPWTQRFFESFGDLSTPDAVMGNPKVKAH
  GKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
  HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQA
  AYQKVVAGVANALAHKYH

BLAST example - Input



Search

```
>sp|P68871|HBB_HUMAN Hemoglobin beta chain - Homo sapiens
(Human).
VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV
KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK
EFTPPVQAAYQKVVAGVANALAHKYH
```

Set subsequence From: To:

Choose database: nr

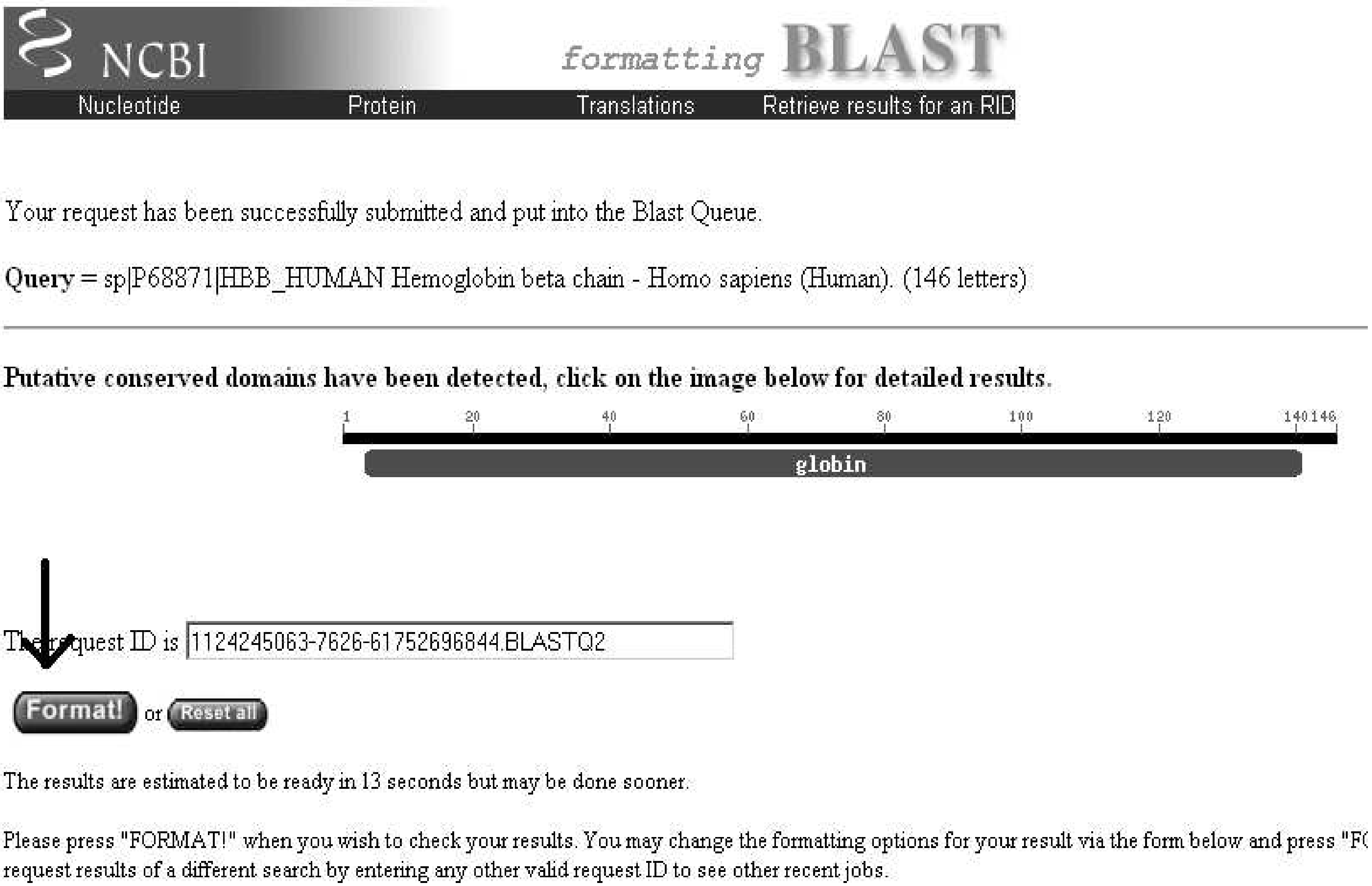
Do CD-Search ☑

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

BLAST example - Output



Your request has been successfully submitted and put into the Blast Queue.

Query = sp|P68871|HBB_HUMAN Hemoglobin beta chain - Homo sapiens (Human). (146 letters)

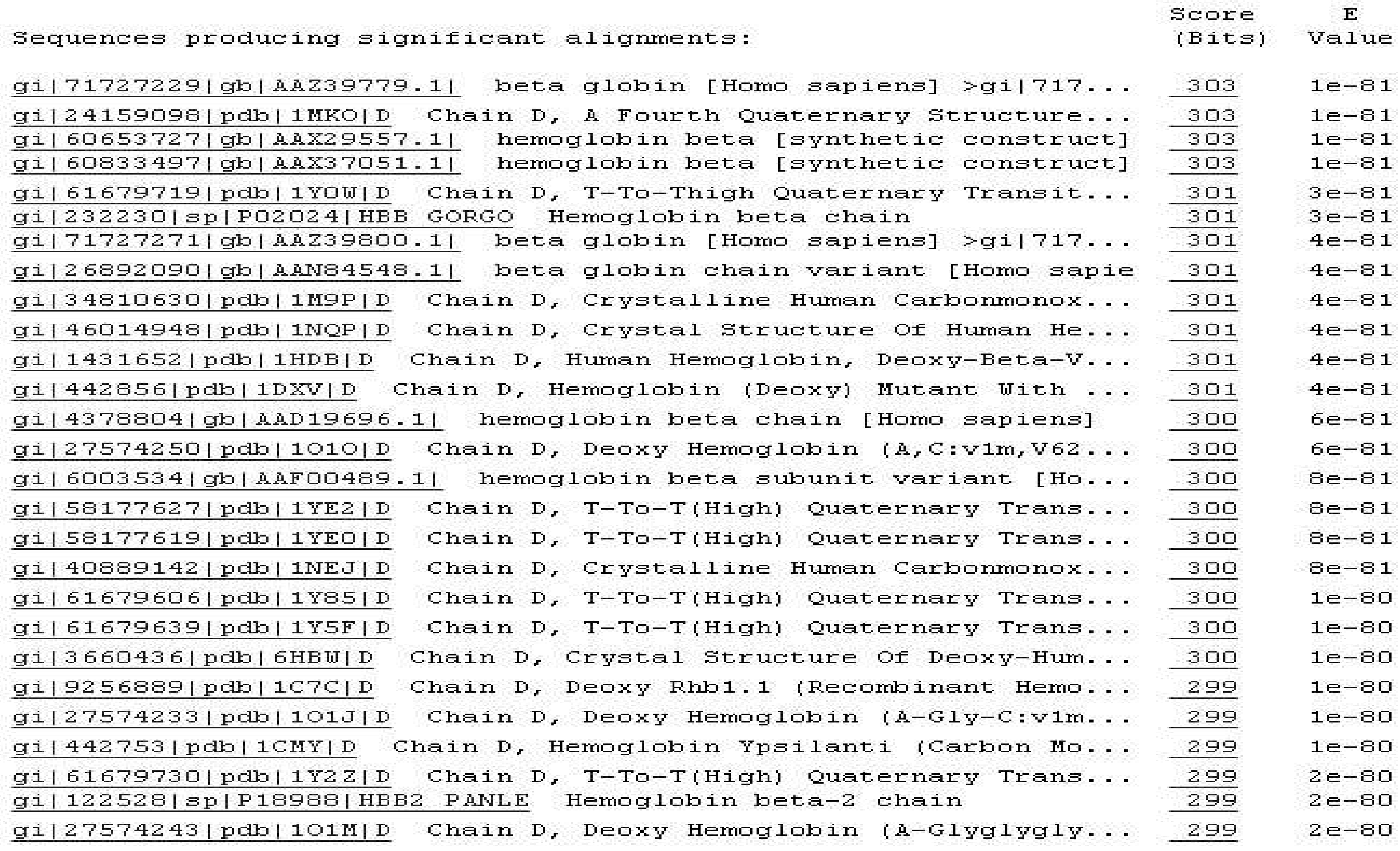
Putative conserved domains have been detected, click on the image below for detailed results.

The request ID is 1124245063-7626-61752696844.BLASTQ2

The results are estimated to be ready in 13 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "F
request results of a different search by entering any other valid request ID to see other recent jobs.

BLAST example - Output



```
                                                                 Score    E
Sequences producing significant alignments:                      (Bits)  Value

gi|71727229|gb|AAZ39779.1|  beta globin [Homo sapiens] >gi|717...  303   1e-81
gi|24159098|pdb|1MKO|D  Chain D, A Fourth Quaternary Structure...  303   1e-81
gi|60653727|gb|AAX29557.1|  hemoglobin beta [synthetic construct]  303   1e-81
gi|60833497|gb|AAX37051.1|  hemoglobin beta [synthetic construct]  303   1e-81
gi|61679719|pdb|1YOW|D  Chain D, T-To-Thigh Quaternary Transit...  301   3e-81
gi|232230|sp|P02024|HBB_GORGO  Hemoglobin beta chain               301   3e-81
gi|71727271|gb|AAZ39800.1|  beta globin [Homo sapiens] >gi|717...  301   4e-81
gi|26892090|gb|AAN84548.1|  beta globin chain variant [Homo sapie  301   4e-81
gi|34810630|pdb|1M9P|D  Chain D, Crystalline Human Carbonmonox...  301   4e-81
gi|46014948|pdb|1NQP|D  Chain D, Crystal Structure Of Human He...  301   4e-81
gi|1431652|pdb|1HDB|D  Chain D, Human Hemoglobin, Deoxy-Beta-V...  301   4e-81
gi|442856|pdb|1DXV|D  Chain D, Hemoglobin (Deoxy) Mutant With ...  301   4e-81
gi|4378804|gb|AAD19696.1|  hemoglobin beta chain [Homo sapiens]    300   6e-81
gi|27574250|pdb|1O1O|D  Chain D, Deoxy Hemoglobin (A,C:v1m,V62...  300   6e-81
gi|6003534|gb|AAF00489.1|  hemoglobin beta subunit variant [Ho...  300   8e-81
gi|58177627|pdb|1YE2|D  Chain D, T-To-T(High) Quaternary Trans...  300   8e-81
gi|58177619|pdb|1YE0|D  Chain D, T-To-T(High) Quaternary Trans...  300   8e-81
gi|40889142|pdb|1NEJ|D  Chain D, Crystalline Human Carbonmonox...  300   8e-81
gi|61679606|pdb|1Y85|D  Chain D, T-To-T(High) Quaternary Trans...  300   1e-80
gi|61679639|pdb|1Y5F|D  Chain D, T-To-T(High) Quaternary Trans...  300   1e-80
gi|3660436|pdb|6HBW|D  Chain D, Crystal Structure Of Deoxy-Hum...  300   1e-80
gi|9256889|pdb|1C7C|D  Chain D, Deoxy Rhb1.1 (Recombinant Hemo...  299   1e-80
gi|27574233|pdb|1O1J|D  Chain D, Deoxy Hemoglobin (A-Gly-C:v1m...  299   1e-80
gi|442753|pdb|1CMY|D  Chain D, Hemoglobin Ypsilanti (Carbon Mo...  299   1e-80
gi|61679730|pdb|1Y2Z|D  Chain D, T-To-T(High) Quaternary Trans...  299   2e-80
gi|122528|sp|P18988|HBB2_PANLE  Hemoglobin beta-2 chain            299   2e-80
gi|27574243|pdb|1O1M|D  Chain D, Deoxy Hemoglobin (A-Glyglygly...  299   2e-80
```

References

- http://www.ncbi.nlm.nih.gov/