Barber notes

The database is the thing...using 'nr' will pull out lots of hits that look like pretty much the same thing (for example, there might be 14 strains of E. coli that contain basically the same protein that all show up)...using more refined databases like 'refseq_select or refseq_protein' identify more variants because fewer sequences are represented in them. What we have found is database choice is influenced by the sequence you are looking for....for example, the attached poster has a table using the GSH-FDH example in the signature file...this is a highly conserved protein. the SmtB homolog, i think, is fine with 'nr' though regarding identifying variants....I suggest testing the three databases in the table.

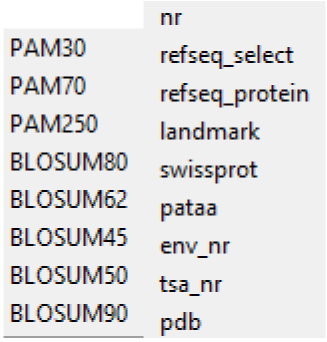
Info on running nev blast

NEVBlastMainGUI.py = the graphical user interface to submit data

NEVBlastBlast.py – obtains initial protein matches to submission

Input parameter	Input description
Matrix	contains the matrices NCBI uses for scoring
Database	selection of various subsets of sequences
Sequence	either the NCBI accessions number or the amino acid sequence to be scored
E-value	maximum e-value allowed from the BLAST
Number of Hits (Hitlist)	the maximum number of sequences to be returned
File	name of the file storing the BLAST report
Organism	optional condition causing the BLAST report to only contain proteins from that organism

Matrix: database (nr = NCBI blast database)



PAM30
PAM70
PAM250
BLOSUM80
BLOSUM62
BLOSUM45
BLOSUM50
BLOSUM90

nr
refseq_select
refseq_protein
landmark
swissprot
pataa
env_nr
tsa_nr
pdb

PAM = global alignment – sequence ‘end to end alignment’ (good for closely related proteins)

PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence but corresponds to 99% sequence identity.

BLOSUM = local alignment – only regions of high degree of similarity are aligned

BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.

*I have only used BLOSUM62, Dr Barber said there isn’t much difference between the matrices from what other students have looked at, in theory PAM should be quite different then BLOSUM.

Data in the blast window is submitted through the NEVBlastBlast.py, this is an API (application program interface) based submission (lines that contain NCBIWWW) obtains alignment data in an xml file format.

NEVBlastBackend.py

Performs a secondary BLOSUM62 alignment of submitted and blast hit sequences, this is where the signatures are searched for.

See last page of signature search submission

Plot.py

Uses the submitted

Methanosarcina acetivorans SmtB homolog

>AAM07687.1 efflux system transcriptional regulator, ArsR family [Methanosarcina acetivorans C2A]

Sequence:

MQEKCDRVNPEQIENLLQKVPDPEYITRMSAVFQALQSDTRLKILFLLRQKEMCVCELEQALEVTQSAVS
HGLRTLRQLDLVRVRREGKFTVYYIADEHVRTLIEMCLEHVEEKI

Signature for quins – use Blosum62

Name of submission for initial blast

[C5, C54, C56], [S67, S70, H71, L76, Y93]

Plot

Name of submission for signature file

Signature 1:

C 5 C 54 C 56

Signature 2:

S 67 S 70 H 71 L 76 Y 93

Input parameter	Input description
Matrix	BLOSUM62
Database	Test the 3: nr, refseq_select, refseq_protein
Sequence	AAM07687.1 trying: AAO75904.1
E-value	0.0001
Number of Hits (Hitlist)	100
File	SmtBhprac
Organism	(leave empty)

Color: chosen from poster.