# March 4, 2013

## Making the autotransporter alignment

All this was done using Python 2.72, IPython 0.13, Biopython 1.57, and ClustalΩ 1.1.0.

### With the Gonnet matrix

I just truncated the gonnet matrix alignment I already had in "family moments/gonnet aligned" using the following script:

**import** Bio**.**AlignIO

msa **=** Bio**.**AlignIO**.**read**(**'gonnet aligned/1UYN with 12.1.6.clu'**,** 'clustal'**)**

template\_seq **=** msa**[-**3**]**

first\_col **=** template\_seq**.**seq**.**find**(**'AATVYA'**)**

**assert** first\_col **==** 1893

trunc\_msa **=** msa**[:,**1893**:]**

Bio**.**AlignIO**.**write**(**trunc\_msa**,** 'truncated 1UYN with 12.1.6.fasta'**,** 'fasta'**)**

See that folder's log for a history of the file.

I sent the resulting file "truncated 1UYN with 12.1.6.fasta" to Sagar Khare and Vik.

### With ClustalOmega

Ran ClustalΩ using the following command:

C:\cygwin\home\alex\beta-barrel-oligomerization\alignment for sagar>clustalo -i

"1UYN with 12.1.6.clu" --dealign -o "better 1UYN with 12.1.6.clu" -v > clustalol

og.txt

It gave the following messages (with progress bars removed):

Using 8 threads

Read 58 sequences (type: Protein) from 1UYN with 12.1.6.clu

Dealigning already aligned input sequences as requested.

Using 34 seeds (chosen with constant stride from length sorted seqs) for mBed (from a total of 58 sequences)

Calculating pairwise ktuple-distances...

Ktuple-distance calculation progress: 0 % (0 out of 1411)*etc...*

Ktuple-distance calculation progress: 95 % (1353 out of 1411)Ktuple-distance calculation progress done. CPU time: 0.57u 00:00:00.57 Elapsed: 00:00:00

mBed created 1 cluster/s (with a minimum of 1 and a soft maximum of 100 sequences each)

Distance calculation within sub-clusters: 0 % (0 out of 1)Distance calculation within sub-clusters done. CPU time: 0.48u 00:00:00.47 Elapsed: 00:00:01

Guide-tree computation (mBed) done.

Progressive alignment progress: 1 % (1 out of 57)*etc...* Progressive alignment progress: 100 % (57 out of 57)Progressive alignment progress done. CPU time: 23.55u 00:00:23.54 Elapsed: 00:00:23

Alignment written to better 1UYN with 12.1.6.clu

However, I could not open the resulting file in Biopython as a clustal file. It seems to actually be in FASTA format. I truncated it with the following script:

**import** Bio**.**AlignIO

msa **=** Bio**.**AlignIO**.**read**(**'better 1UYN with 12.1.6.clu'**,** 'fasta'**)**

template\_seq **=** msa**[-**3**]**

first\_col **=** template\_seq**.**seq**.**find**(**'AATVYA'**)**

trunc\_msa **=** msa**[:,**first\_col**:]**

Bio**.**AlignIO**.**write**(**trunc\_msa**,** 'better truncated 1UYN with 12.1.6.fasta'**,** 'fasta'**)**

I then checked the sequence identities:

In [72]: from seqtools import identity

In [73]: id\_ = identity(trunc\_msa[-3], trun

trunc\_msa truncated\ 1UYN\ with\ 12.1.6.fasta

In [73]: id\_ = identity(trunc\_msa[-3], trunc\_msa)

In [74]: id\_

Out[74]:

[('gi|67547523|ref|ZP', 0.06027397260273973),

('gi|543788|sp|Q0315', 0.0782122905027933),

('gi|77972002|ref|ZP', 0.08868501529051988),

('gi|16122995|ref|NP', 0.09174311926605505),

('gi|77975979|ref|ZP', 0.09422492401215805),

('gi|16123205|ref|NP', 0.0972644376899696),

('gi|77957624|ref|ZP', 0.0972644376899696),

('gi|77975760|ref|ZP', 0.0972644376899696),

('gi|7466262|pir||G6', 0.09803921568627451),

('gi|82776499|ref|YP', 0.10030395136778116),

('gi|49475344|ref|YP', 0.10060975609756098),

('gi|33599810|ref|NP', 0.10119047619047619),

('gi|15832769|ref|NP', 0.1033434650455927),

('gi|77961204|ref|ZP', 0.1033434650455927),

('gi|26247147|ref|NP', 0.10364145658263306),

('gi|51597164|ref|YP', 0.10429447852760736),

('gi|15800732|ref|NP', 0.10644257703081232),

('gi|28869427|ref|NP', 0.10703363914373089),

('gi|16119591|ref|NP', 0.10714285714285714),

('gi|77963418|ref|ZP', 0.10784313725490197),

('gi|73537641|ref|YP', 0.10914454277286136),

('gi|47154998|emb|CA', 0.1092436974789916),

('gi|66045274|ref|YP', 0.11009174311926606),

('gi|77977360|ref|ZP', 0.11042944785276074),

('gi|113876548|ref|Z', 0.11078717201166181),

('gi|22034298|gb|AAL', 0.11280487804878049),

('gi|83751177|ref|ZP', 0.11314984709480122),

('gi|77459122|ref|YP', 0.11314984709480122),

('gi|33602087|ref|NP', 0.11314984709480122),

('gi|33596254|ref|NP', 0.11314984709480122),

('gi|33602086|ref|NP', 0.11314984709480122),

('gi|77974466|ref|ZP', 0.11349693251533742),

('gi|104782218|ref|Y', 0.11538461538461539),

('gi|26989788|ref|NP', 0.11538461538461539),

('gi|83751176|ref|ZP', 0.11585365853658537),

('gi|984283|gb|AAC43', 0.1162079510703364),

('gi|465619|sp|P3392', 0.11854103343465046),

('gi|49474105|ref|YP', 0.11890243902439024),

('gi|26988610|ref|NP', 0.11926605504587157),

('gi|71736402|ref|YP', 0.11926605504587157),

('gi|66046637|ref|YP', 0.11926605504587157),

('gi|33602267|ref|NP', 0.11926605504587157),

('gi|77974678|ref|ZP', 0.1196319018404908),

('gi|104783283|ref|Y', 0.12158054711246201),

('gi|28869215|ref|NP', 0.12232415902140673),

('gi|77957189|ref|ZP', 0.12576687116564417),

('gi|51597166|ref|YP', 0.12883435582822086),

('gi|49475345|ref|YP', 0.13109756097560976),

('gi|49475343|ref|YP', 0.13109756097560976),

('gi|75474399|sp|Q9X', 0.13414634146341464),

('gi|33593700|ref|NP', 0.13761467889908258),

('gi|34497333|ref|NP', 0.14067278287461774),

('gi|82736677|ref|ZP', 0.1437308868501529),

('gi|33595824|ref|NP', 0.1437308868501529),

('gi|33601253|ref|NP', 0.14678899082568808),

('gi|33601917|ref|NP', 0.15902140672782875),

('gi|33594369|ref|NP', 0.1620795107033639),

('template\_1UYN', 1.0)]

I saved this list to a CSV file, backwards so nobody would take an average including the template by mistake:

In [78]: csv.writer(open('autotransporter sequence identities.csv', 'wb')).write

rows(id\_[::-1])