Room E3607
Protein bioinformatics
260.841 Protein Bioinformatics

Computer lab
Tuesday, May 17, 2005
Sean Prigge
Jonathan Pevsner
Ingo Ruczinski

Outline of today's lab

	Topic	Suggested time
1	Find a protein's primary amino acid sequence	15 minutes
2	Assess the secondary structure of the protein	15 minutes
3	Find out whether three-dimensional structure data are available for this protein (or a region of the protein)	15 minutes
4	Analyze the 3D structure of a protein	30 minutes
5	2D gels, mass spec data, and protein interaction networks	as available

We will focus on six proteins that you told us you are currently researching:

	Protein
1	Yeast Esa1
2	Human IL-13
3	Rat tech: neuronal rho gef
4	Human huntingtin
5	Nuclear receptor NR2E3
6	CD4-like Lag3

Part 1 (suggested time: 15 minutes)
Topic: find the right protein sequence
Approach 1: NCBI
Approach 2: ExPASy

	Protein (by name)	Entrez Gene result (NCBI)		
		#	Name	RefSeq or Gb
1	Yeast Esa1	7	ESA1 Esa1p [Saccharomyces cerevisiae] Other Aliases: YOR244W, TAS1	NP_014887
2	Human IL-13	83	IL13 interleukin 13 [Homo sapiens]	NP_998968
3	Rat tech: neuronal rho gef	2	RGD:1303132 Interim Symbol: Tech and Name: neuronal RhoA GEF protein [Rattus norvegicus]	NP_958429
4	Human huntingtin	214 (172 H.s.)	huntingtin [Homo sapiens]	NP_002102
5	Nuclear receptor NR2E3	7	NR2E3 nuclear receptor subfamily 2, group E, member 3 [Homo sapiens]	NP_055064 (410 aa) NP_057430 (367 aa)
6	CD4-like Lag3	3	LAG3 lymphocyte-activation gene 3 [Homo sapiens]	NP_002277

Key points:
[1] NCBI Entrez is a good starting point for finding protein and DNA sequences. From the main page of NCBI, enter a search. Then try "Entrez Gene."
[2] Names can be ambiguous, so be careful. For Esa1 there are matches to non-yeast proteins.
[3] For huntingtin there are 214 entries in Entrez Gene. You can reduce the list of results by selecting *Home sapiens*, and by using the limit option to select "from RefSeq only."
[4] RefSeq (reference sequence) accession numbers have the format NP_xxxxxx (for proteins) or NM_xxxxxx (for DNA). RefSeq is preferred because these are curated (agree-upon by experts).

From the Entrez Protein page, select “FASTA” from the display options for each protein. The FASTA format consists of a carat > followed by an unbroken string of text (it can wrap across several lines, as shown in the first example below). The FASTA format is extremely useful for pasting your protein sequence into a variety of on-line tools, such as secondary structure prediction programs.

```
>gi|6324818|ref|NP_014887.1| Histone acetyltransferase catalytic
subunit of the native multisubunit complex (NuA4) that acetylates four,
conserved internal lysines of histone H4 N-terminal tail; required for
cell cycle progression; Esa1p [Saccharomyces cerevisiae]
MSHDGKEEPGIAKKINSVDDIIIKCQCWVQKNDEERLAEILSINTRKAPPKFYVHYVNYNKRLDEWITTD
RINLDKEVLYPKLKATDEDNKKQKKKKATNTSETPQDSLQDGVDGFSRENTDVMDLDNLNVQGIKDENIS
HEDEIKKLRTSGSMTQNPHEVARVRNLNRIIMGKYEIEPWYFSPYPIELTDEDFIYIDDFTLQYFGSKKQ
YERYRKKCTLRHPPGNEIYRDDYVSFFEIDGRKQRTWCRNLCLLSKLFLDHKTLYYDVDPFLFYCMTRRD
ELGHHLVGYFSKEKESADGYNVACILTLPQYQRMGYGKLLIEFSYELSKKENKVGSPEKPLSDLGLLSYR
AYWSDTLITLLVEHQKEITIDEISSMTSMTTTDILHTAKTLNILRYYKGQHIIFLNEDILDRYNRLKAKK
RRTIDPNRLIWKPPVFTASQLRFAW

>gi|47523238|ref|NP_998968.1| interleukin-13 [Sus scrofa]
MALWLTLVIALTCFGGLASPGPVPPHSTALKELIEELVNITQNQKTPLCNGSMVWSVNLTTSMQYCAALE
SLINISDCSAIQKTQRMLSALCSHKPPSEQVPGKHIRDTKIEVAQFVKDLLKHLRMIFRHG

>gi|41152499|ref|NP_958429.1| neuronal RhoA GEF protein [Rattus
norvegicus]
MDKGRAAKVCHHADCQQLHHRGPLNLCEICDSKFHNTTHYDGHVRFDLPPQGSVLARNVSTRSCPPRTSP
AGDLEEEDEGYTNGKGDRKSAGLKISKKKARRRHTDDPSKECFTLKFDLNVDIETEIVPAMKKKSLGEVL
LPVFERKGIALGKVDIYLDQSNTPLSLTFEAYRFGGHYLRVKAKPGDEGKVEQGVKDSKSLSLPALRPSG
AGTPVLERVDPQSRRESSLDILAPGRRRKNMSEFLGDTSIPGQESPAPSSCSLPVGSSVGSSGSSESWKN
RAASRFSGFFSSSPSTGAFGREVDKMEQLESKLHAYSLFGLPRMPRRLRFDHDSWEEEEDDEEEEDNSGL
RLEDSWRELIDGHEKLTRRQCHQQEAVWELLHTEVSYIRKLRVITNLFLCCLLNLQESGLLCEVEAERLF
SNIPELARLHRGLWSSVMVPVLEKARRTRALLQPSDFLKGFKMFGSLFKPYIRYCMEEEGCMEYMRSLLR
DNDLFRAYVTWAEKHQQCQRLKLSDMLAKPHQRLTKYPLLLKSVLRKTDEPRAKEAIITMISSVERFIHH
VNTCMRQRQERQRLAGVVSRIDAYEVVEGSNDEVDKFLKEFLHLDLTAPMPGTSPEETRQLLLEGSLRMK
EGKDSKMDVYCFLFTDLLLVTKAVKKAERTKVIRPPLLVDKIVCRELRDPGSFLLIHLNEFHSAVGAYTF
QASSQALCRSWVDTLYNAQNQLQQLRAQLLCAQEHPGTQHLQSLEEEEDEQEEEGEESGTSAASSPTILR
KSSNSLDSEHCASDGSTETLAMVVVEPGETLSSPEFDRGPFSSQSDEASLSNTTSSITPTSELLPLGPVD
GRSCSMDSAYGTLSPTSLQDFAAPHPVVEPVPVPQTLSPQPSPRLRRRTPVQLLPRLPHLLKSKSEASLL
QLLSGTTTSVSPPAPSRSLSELCLITMAPGVRTQSSLQEGGPGWNCPGACGPCQGPPLSESENRPSHKAG
GPADSARRKCREMPCGTVPRVQPEPSPGISAQHRKLTLAQLYRIRTTLLLNSTLTASEV

>gi|4753163|ref|NP_002102.2| huntingtin [Homo sapiens]
MATLEKLMKAFESLKSFQQQQQQQQQQQQQQQQQQQQQQQPPPPPPPPPPQLPQPPPQAQPLLPQPQPP
PPPPPPPPGPAVAEEPLHRPKKELSATKKDRVNHCLTICENIVAQSVRNSPEFQKLLGIAMELFLLCSDD
AESDVRMVADECLNKVIKALMDSNLPRLQLELYKEIKKNGAPRSLRAALWRFAELAHLVRPQKCRPYLVN
LLPCLTRTSKRPEESVQETLAAAVPKIMASFGNFANDNEIKVLLKAFIANLKSSSPTIRRTAAGSAVSIC
QHSRRTQYFYSWLLNVLLGLLVPVEDEHSTLLILGVLLTLRYLVPLLQQQVKDTSLKGSFGVTRKEMEVS
PSAEQLVQVYELTLHHTQHQDHNVVTGALELLQQLFRTPPPELLQTLTAVGGIGQLTAAKEESGGRSRSG
SIVELIAGGGSSCSPVLSRKQKGKVLLGEEEALEDDSESRSDVSSSALTASVKDEISGELAASSGVSTPG
SAGHDIITEQPRSQHTLQADSVDLASCDLTSSATDGDEEDILSHSSSQVSAVPSDPAMDLNDGTQASSPI
SDSSQTTTEGPDSAVTPSDSSEIVLDGTDNQYLGLQIGQPQDEDEEATGILPDEASEAFRNSSMALQQAH
LLKNMSHCRQPSDSSVDKFVLRDEATEPGDQENKPCRIKGDIGQSTDDDSAPLVHCVRLLSASFLLTGGK
NVLVPDRDVRVSVKALALSCVGAAVALHPESFFSKLYKVPLDTTEYPEEQYVSDILNYIDHGDPQVRGAT
AILCGTLICSILSRSRFHVGDWMGTIRTLTGNTFSLADCIPLLRKTLKDESSVTCKLACTAVRNCVMSLC
SSSYSELGLQLIIDVLTLRNSSYWLVRTELLETLAEIDFRLVSFLEAKAENLHRGAHHYTGLLKLQERVL
NNVVIHLLGDEDPRVRHVAAASLIRLVPKLFYKCDQGQADPVVAVARDQSSVYLKLLMHETQPPSHFSVS
```

TITRIYRGYNLLPSITDVTMENNLSRVIAAVSHELITSTTRALTFGCCEALCLLSTAFPVCIWSLGWHCG
VPPLSASDESRKSCTVGMATMILTLLSSAWFPLDLSAHQDALILAGNLLAASAPKSLRSSWASEEEANPA
ATKQEEVWPALGDRALVPMVEQLFSHLLKVINICAHVLDDVAPGPAIKAALPSLTNPPSLSPIRRKGKEK
EPGEQASVPLSPKKGSEASAASRQSDTSGPVTTSKSSSLGSFYHLPSYLKLHDVLKATHANYKVTLDLQN
STEKFGGFLRSALDVLSQILELATLQDIGKCVEEILGYLKSCFSREPMMATVCVQQLLKTLFGTNLASQF
DGLSSNPSKSQGRAQRLGSSSVRPGLYHYCFMAPYTHFTQALADASLRNMVQAEQENDTSGWFDVLQKVS
TQLKTNLTSVTKNRADKNAIHNHIRLFEPLVIKALKQYTTTTCVQLQKQVLDLLAQLVQLRVNYCLLDSD
QVFIGFVLKQFEYIEVGQFRESEAIIPNIFFFLVLLSYERYHSKQIIGIPKIIQLCDGIMASGRKAVTHA
IPALQPIVHDLFVLRGTNKADAGKELETQKEVVVSMLLRLIQYHQVLEMFILVLQQCHKENEDKWKRLSR
QIADIILPMLAKQQMHIDSHEALGVLNTLFEILAPSSLRPVDMLLRSMFVTPNTMASVSTVQLWISGILA
ILRVLISQSTEDIVLSRIQELSFSPYLISCTVINRLRDGDSTSTLEEHSEGKQIKNLPEETFSRFLLQLV
GILLEDIVTKQLKVEMSEQQHTFYCQELGTLLMCLIHIFKSGMFRRITAAATRLFRSDGCGGSFYTLDSL
NLRARSMITTHPALVLLWCQILLLVNHTDYRWWAEVQQTPKRHSLSSTKLLSPQMSGEEEDSDLAAKLGM
CNREIVRRGALILFCDYVCQNLHDSEHLTWLIVNHIQDLISLSHEPPVQDFISAVHRNSAASGLFIQAIQ
SRCENLSTPTMLKKTLQCLEGIHLSQSGAVLTLYVDRLLCTPFRVLARMVDILACRRVEMLLAANLQSSM
AQLPMEELNRIQEYLQSSGLAQRHQRLYSLLDRFRLSTMQDSLSPSPPVSSHPLDGDGHVSLETVSPDKD
WYVHLVKSQCWTRSDSALLEGAELVNRIPAEDMNAFMMNSEFNLSLLAPCLSLGMSEISGGQKSALFEAA
REVTLARVSGTVQQLPAVHHVFQPELPAEPAAYWSKLNDLFGDAALYQSLPTLARALAQYLVVVSKLPSH
LHLPPEKEKDIVKFVVATLEALSWHLIHEQIPLSLDLQAGLDCCCLALQLPGLWSVVSSTEFVTHACSLI
YCVHFILEAVAVQPGEQLLSPERRTNTPKAISEEEEEVDPNTQNPKYITAACEMVAEMVESLQSVLALGH
KRNSGVPAFLTPLLRNIIISLARLPLVNSYTRVPPLVWKLGWSPKPGGDFGTAFPEIPVEFLQEKEVFKE
FIYRINTLGWTSRTQFEETWATLLGVLVTQPLVMEQEESPPEEDTERTQINVLAVQAITSLVLSAMTVPV
AGNPAVSCLEQQPRNKPLKALDTRFGRKLSIIRGIVEQEIQAMVSKRENIATHHLYQAWDPVPSLSPATT
GALISHEKLLLQINPERELGSMSYKLGQVSIHSVWLGNSITPLREEEWDEEEEEEADAPAPSSPPTSPVN
SRKHRAGVDIHSCSQFLLELYSRWILPSSSARRTPAILISEVVRSLLVVSDLFTERNQFELMYVTLTELR
RVHPSEDEILAQYLVPATCKAAAVLGMDKAVAEPVSRLLESTLRSSHLPSRVGALHGVLYVLECDLLDDT
AKQLIPVISDYLLSNLKGIAHCVNIHSQQHVLVMCATAFYLIENYPLDVGPEFSASIIQMCGVMLSGSEE
STPSIIYHCALRGLERLLLSEQLSRLDAESLVKLSVDRVNVHSPHRAMAALGLMLTCMYTGKEKVSPGRT
SDPNPAAPDSESVIVAMERVSVLFDRIRKGFPCEARVVARILPQFLDDFFPPQDIMNKVIGEFLSNQQPY
PQFMATVVYKVFQTLHSTGQSSMVRDWVMLSLSNFTQRAPVAMATWSLSCFFVSASTSPWVAAILPHVIS
RMGKLEQVDVNLFCLVATDFYRHQIEEELDRRAFQSVLEVVAAPGSPYHRLLTCLRNVHKVTTC

>gi|7657395|ref|NP_055064.1| photoreceptor-specific nuclear receptor isoform b [Homo sapiens]
METRPTALMSSTVAAAAPAAGAASRKESPGRWGLGEDPTGVSPSLQCRVCGDSSSGKHYGIYACNGCSGF
FKRSVRRRLIYRCQVGAGMCPVDKAHRNQCQACRLKKCLQAGMNQDAVQNERQPRSTAQVHLDSMESNTE
SRPESLVAPPAPAGRSPRGPTPMSAARALGHHFMASLITAETCAKLEPEDADENIDVTSNDPEFPSSPYS
SSSPCGLDSIHETSARLLFMAVKWAKNLPVFSSLPFRDQVILLEEAWSELFLLGAIQWSLPLDSCPLLAP
PEASAAGGAQGRLTLASMETRVLQETISRFRALAVDPTEFACMKALVLFKPETRGLKDPEHVEALQDQSQ
VMLSQHSKAHHPSQPVRFGKLLLLLPSLRFITAERIELLFFRKTIGNTPMEKLLCDMFKN

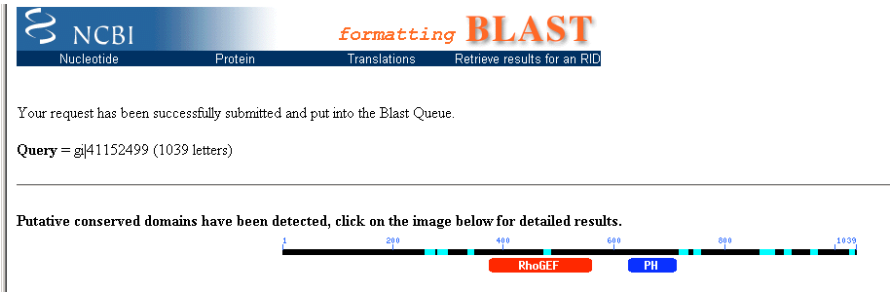
>gi|15718682|ref|NP_002277.3| lymphocyte-activation protein 3 precursor [Homo sapiens]
MWEAQFLGLLFLQPLWVAPVKPLQPGAEVPVVWAQEGAPAQLPCSPTIPLQDLSLLRRAGVTWQHQPDSG
PPAAAPGHPLAPGPHPAAPSSWGPRPRRYTVLSVGPGGLRSGRLPLQPRVQLDERGRQRGDFSLWLRPAR
RADAGEYRAAVHLRDRALSCRLRLRLGQASMTASPPGSLRASDWVILNCSFSRPDRPASVHWFRNRGQGR
VPVRESPHHHLAESFLFLPQVSPMDSGPWGCILTYRDGFNVSIMYNLTVLGLEPPTPLTVYAGAGSRVGL
PCRLPAGVGTRSFLTAKWTPPGGGPDLLVTGDNGDFTLRLEDVSQAQAGTYTCHIHLQEQQLNATVTLAI
ITVTPKSFGSPGSLGKLLCEVTPVSGQERFVWSSLDTPSQRSFSGPWLEAQEAQLLSQPWQCQLYQGERL
LGAAVYFTELSSPGAQRSGRAPGALPAGHLLLFLTLGVLSLLLLVTGAFGFHLWRRQWRPRRFSALEQGI
HPPQAQSKIEELEQEPEPEPEPEPEPEPEPEPEQL

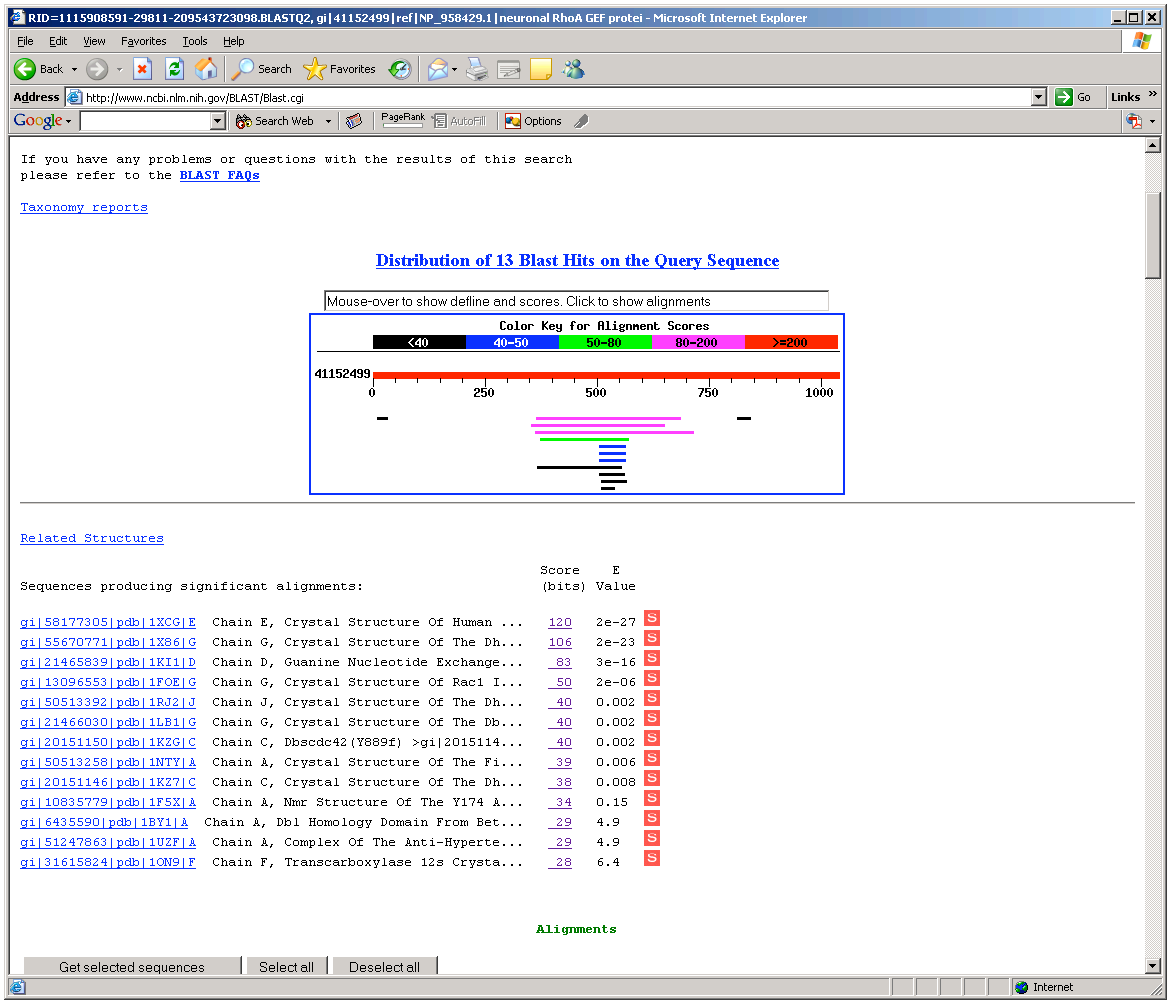
Part 2 (suggested time: 15 minutes)
Topic: assess the secondary structure of the protein
Approach 1: POLE
Approach 2: ExPASy.

Part 3 (suggested time: 15 minutes)
Topic: find out whether three-dimensional structure data are available for this protein (or a region of the protein)
Approach 1: Entrez Structure at NCBI
Approach 2: blastp search at NCBI
Approach 3: text search at PDB
Approach 4: FASTA search at PDB
Example 1: huntingtin

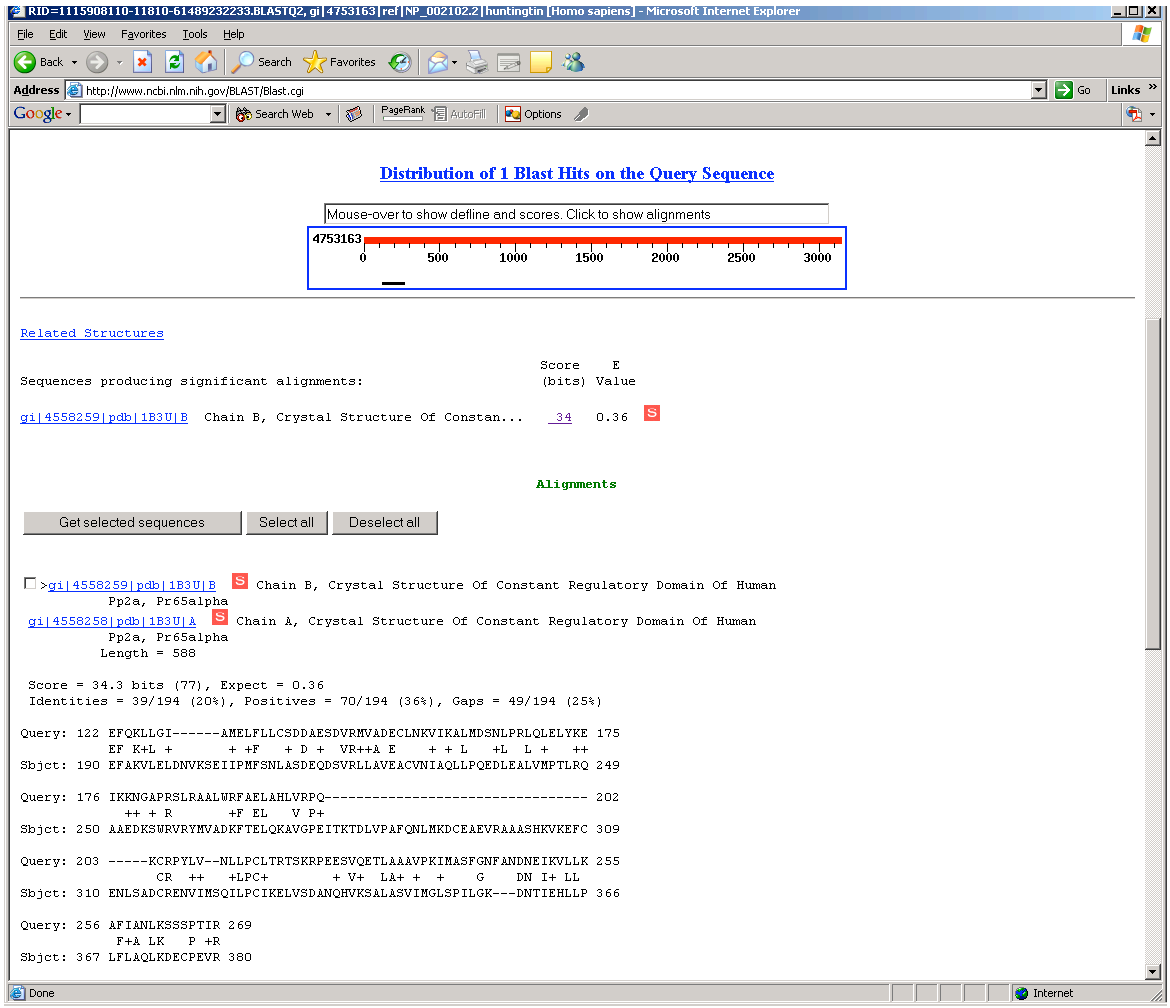
	Entrez Structure	PDB ID	NCBI blast versus PDB
1	4 (ESA1)	1FY7, 1MJB 1MJA, 1MJ9	
	1 (Esa1p)	1FY7	
2	1 (nmr structure)	1GA3	
3	0 matches	---	matches found
4	5 (all HIPs)	---	no sig. matches
5	0 matches	---	matches found
6	0 matches	---	matches found

protein 3: tech NP_958429
blastp versus PDB
matches found e.g. RhoA, 1XCG

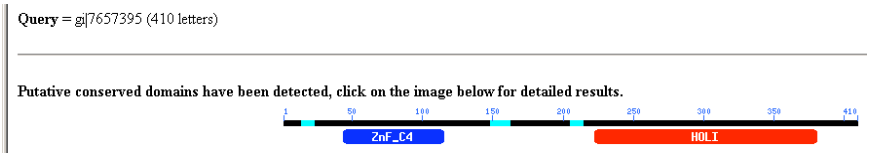


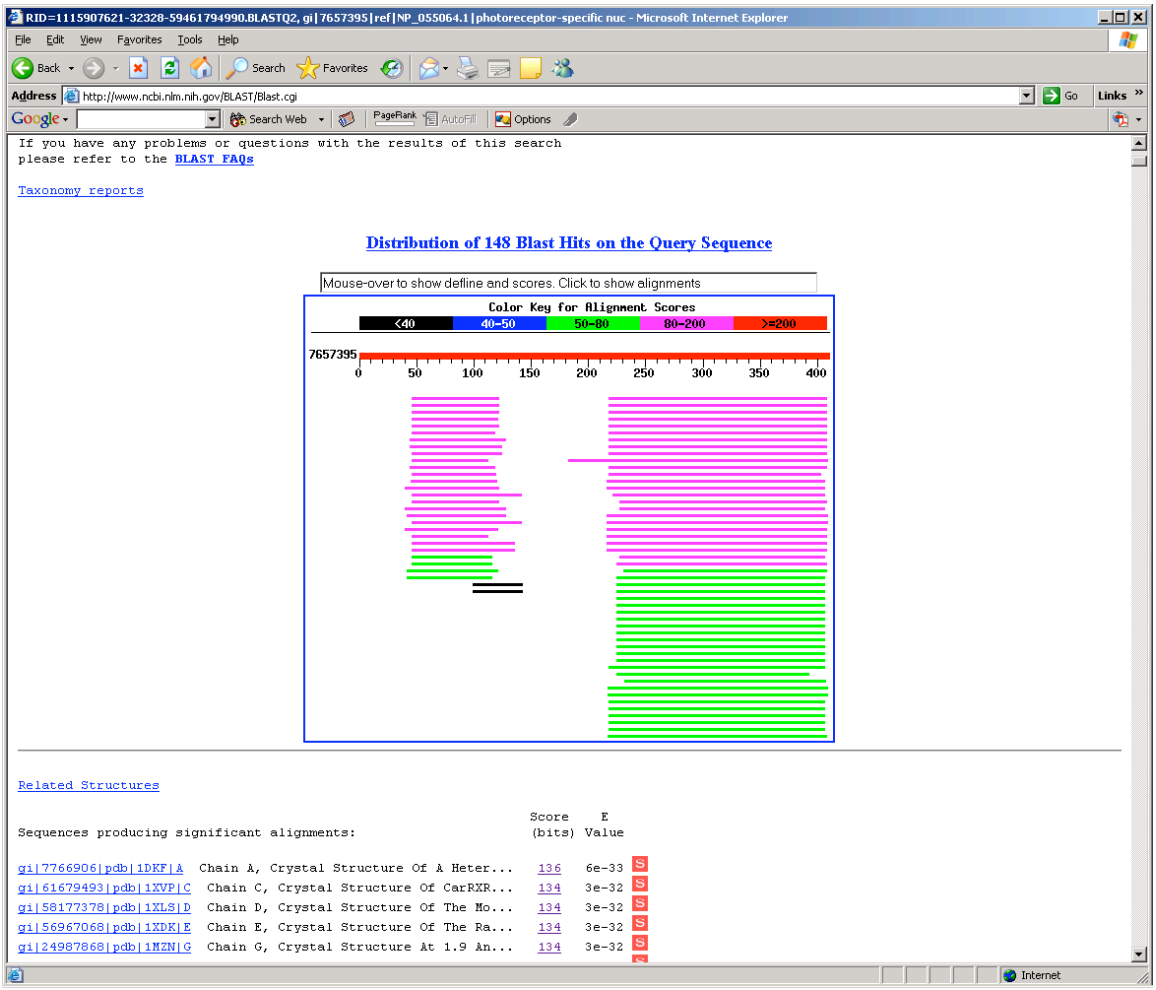


protein 4: huntingtin NP_002102
blastp versus PDB
result: 1 match, not significant



protein 5: NR2E3 NP_055064
blastp versus PDB
result: many matches, e.g. 1DKF (residues 219-409 of NR2E3), 1YNW (vitamin D receptor) (residues 47-123)

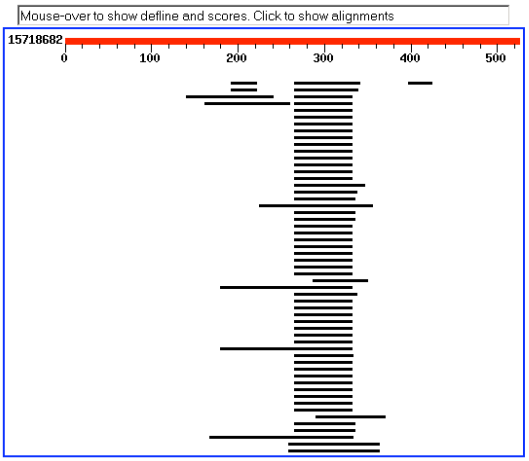




protein 6: lag3 NP_002277
blastp versus pdb
match found, residues 266-342 (approximate)

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

Distribution of 149 Blast Hits on the Query Sequence



Related Structures

Sequences producing significant alignments:		Score (bits)	E Value	
gi|7766934|pdb|32C2|A	Chain A, Structure Of An Activity Sup...	37	0.009	S
gi|1064953|pdb|1IBG|L	Chain L, Igg Fab (Igg2b, Kappa) Fragm...	37	0.009	S
gi|1127091|pdb|1NSN|L	Chain L, Immunoglobulin, Staphylococc...	37	0.011	S
gi|15825780|pdb|1I7Z|C	Chain C, Antibody Gnc92h2 Bound To L...	36	0.019	S
gi|24158825|pdb|1L7T|L	Chain L, Crystal Structure Analysis ...	33	0.096	S
gi|10835827|pdb|1F3D|J	Chain J, Catalytic Antibody 4b2 In C...	33	0.16	S
gi|20150130|pdb|1I9R|Y	Chain Y, Structure Of Cd40l In Compl...	32	0.21	S
gi|4558340|pdb|3F58|L	Chain L, Igg1 Fab Fragment (58.2) Com...	32	0.21	S
gi|4558336|pdb|2F58|L	Chain L, Igg1 Fab Fragment (58.2) Com...	32	0.21	S

Part 4 (suggested time: 30 minutes)
Topic: analyze the 3D structure of a protein
Approach: CN3D

Part 5 (as time permits)
Topic: 2D gels, mass spec data, and protein interaction networks
Approach: visit ExPASy