**Bioinformatic assignment: multiple sequence alignment**

1. You want to find the DNA-binding domain (DBD) in a certain TF (Gln3). DNA binding domains should be conserved across evolution. Use a MSA alignment tool to align the 6 provided TF sequences.
   1. What is the number of identical/similar positions in the alignment?
   2. Where would do you locate the DBD is in Gln3\_Scer?
   3. What other aspect suggests it is a DNA binding domain?
   4. Imagine the following three AA substitutions in Gln3\_Scer, order them by their probability to affect protein function : T89Q, K351R, S344Q
2. One way to estimate the reliability of a multiple sequence alignment of a protein is to compare it against the known 3D structure.
   1. Look a the structure of *E. coli* OmpA in alpha fold (UniProtID: P0A910).
   2. Now align *E. coli* OmpA and eight orthologs to define conserved regions. Which regions of P0A910 are least conserved between the orthologs? GIive the starting and end residues.
   3. Which structural elements (loop, helix, strand) are associated with this region?
   4. Do you expect this? Why?
3. Orthologs and paralogs are homologous proteins, in related species or in the same species respectively. Looking at the Caspase family of cysteine proteases there are several paralogs in humans and multiple orthologs across mammalian species.
   1. Download the AA sequence from 5 different caspase3 (casp3) orthologs in Human, Mouse, Bovine, chicken and xenopus from Uniprot.
   2. Align the 5 ortholog sequences using an MSA algorithm. How many residues are identical, similar?
   3. Download the AA sequence from 5 different caspase paralogs (casp) in Human (Caspase 1,2,3,4 and 5) .
   4. Align the 5 paralog sequences. How many residues are identical, similiar?
   5. Which set of sequences are more similar orthologs or paralogs? What does this suggest about the evolution of caspases in mammals?
   6. You want to find the catalytic cystein residue in caspase 3. Do you expect it to be conserved across orthologs/paralogs?
   7. Which MSA would you use to find it? From the MSA, which residues could be the catalytical cysteine?
4. –
   1. 13 positions
   2. CNACGLFQKLHGTMRPLSLK
5. –
   1. –
   2. 48-56, 96-101, 139-151, 213-222
   3. Strand-loop, loop, loop, loop-helix
6. –
   1. >human MENTENSVDSKSIKNLEPKIIHGSESMDSGISLDNSYKMDYPEMGLCIIINNKNFHKSTGMTSRSGTDVDAANLRETFRNLKYEVRNKNDLTREEIVELMRDVSKEDHSKRSSFVCVLLSHGEEGIIFGTNGPVDLKKITNFFRGDRCRSLTGKPKLFIIQACRGTELDCGIETDSGVDDDMACHKIPVEADFLYAYSTAPGYYSWRNSKDGSWFIQSLCAMLKQYADKLEFMHILTRVNRKVATEFESFSFDATFHAKKQIPCIVSMLTKELYFYH >mouse MENNKTSVDSKSINNFEVKTIHGSKSVDSGIYLDSSYKMDYPEMGICIIINNKNFHKSTGMSSRSGTDVDAANLRETFMGLKYQVRNKNDLTREDILELMDSVSKEDHSKRSSFVCVILSHGDEGVIYGTNGPVELKKLTSFFRGDYCRSLTGKPKLFIIQACRGTELDCGIETDSGTDEEMACQKIPVEADFLYAYSTAPGYYSWRNSKDGSWFIQSLCSMLKLYAHKLEFMHILTRVNRKVATEFESFSLDSTFHAKKQIPCIVSMLTKELYFYH >bovine MENTENSVDSKSIKTSETKILHGSKSMDSGISLEESYKMDYPEMGLCIIINNKNFHENTGMACRSGTDVDAANLRETFMNLKYEVRIKNDLTCKEMLELMSNVSKEDHSKRSSFICVLLSHGEEGIIFGTNGPVNLKKLASFFRGDYCRSLTGKPKLFIIQACRGTELDCGIETDSGAEDDMACQKIPVEADFLYAYSTAPGYFSWRNAKNGSWFIQALCEMLKKHAHRLELMHILTRVNRKVAIEYESFSTDSAFHAKKQIPCIMSMLTKELYF >chicken MMTDIKDGPRSGEDVSDARSFPGSKGMNLPASKSVDSGILPDDSYRMDYPEIGVCVIINNKNFHRDTGLSSRSGTDADAASVREVFMKLGYKVKLNNDLSSRDIFKLLKNVSEEDHSKRSSFVCVLLSHGDEGLFYGTDGPLELKVLTSLFRGDKCRSLAGKPKLFFIQACRGTELDSGIEADSGPDETVCQKIPVEADFLYAYSTAPGYYSWRNAAEGSWFIQSLCRMLKEHARKLELMQILTRVNRRVAEYESCSTRQDFNAKKQIPCIVSMLTKEFYFPC >xenopus MEESQNGVKYGGDATDAKEYFTIQPRSLQNCDLKDIERKTKFAHLQNYRTNYPEMGMCLIINNKNFHSSNMAVRNGTDVDALKLHETFTGLGYEVMVCNDQKSSDIIGRLKKISEEDHSKRSSFVCAILSHGEEDGSICGVDVPIHIKNLTDLFRGDRCKTLVGKPKIFFIQACRGTELDSGIETDSCSEPREEIQRIPVEADFLYAYSTVPGYCSWRDKMDGSWFIQSLCKMIKLYGSHLELIQILTCVNHMVALDFETFHAKKQIPCVVSMLTKSFYFFK
   2. 4 main blocks of similarity
   3. >casp1 MADKVLKEKRKLFIRSMGEGTINGLLDELLQTRVLNKEEMEKVKRENATVMDKTRALIDSVIPKGAQACQICITYICEEDSYLAGTLGLSADQTSGNYLNMQDSQGVLSSFPAPQAVQDNPAMPTSSGSEGNVKLCSLEEAQRIWKQKSAEIYPIMDKSSRTRLALIICNEEFDSIPRRTGAEVDITGMTMLLQNLGYSVDVKKNLTASDMTTELEAFAHRPEHKTSDSTFLVFMSHGIREGICGKKHSEQVPDILQLNAIFNMLNTKNCPSLKDKPKVIIIQACRGDSPGVVWFKDSVGVSGNLSLPTTEEFEDDAIKKAHIEKDFIAFCSSTPDNVSWRHPTMGSVFIGRLIEHMQEYACSCDVEEIFRKVRFSFEQPDGRAQMPTTERVTLTRCFYLFPGH

>casp 2

MAAPSAGSWSTFQHKELMAADRGRRILGVCGMHPHHQETLKKNRVVLAKQLLLSELLEHLLEKDIITLEMRELIQAKVGSFSQNVELLNLLPKRGPQAFDAFCEALRETKQGHLEDMLLTTLSGLQHVLPPLSCDYDLSLPFPVCESCPLYKKLRLSTDTVEHSLDNKDGPVCLQVKPCTPEFYQTHFQLAYRLQSRPRGLALVLSNVHFTGEKELEFRSGGDVDHSTLVTLFKLLGYDVHVLCDQTAQEMQEKLQNFAQLPAHRVTDSCIVALLSHGVEGAIYGVDGKLLQLQEVFQLFDNANCPSLQNKPKMFFIQACRGDETDRGVDQQDGKNHAGSPGCEESDAGKEKLPKMRLPTRSDMICGYACLKGTAAMRNTKRGSWYIEALAQVFSERACDMHVADMLVKVNALIKDREGYAPGTEFHRCKEMSEYCSTLCRHLYLFPGHPPT

>casp3

MENTENSVDSKSIKNLEPKIIHGSESMDSGISLDNSYKMDYPEMGLCIIINNKNFHKSTGMTSRSGTDVDAANLRETFRNLKYEVRNKNDLTREEIVELMRDVSKEDHSKRSSFVCVLLSHGEEGIIFGTNGPVDLKKITNFFRGDRCRSLTGKPKLFIIQACRGTELDCGIETDSGVDDDMACHKIPVEADFLYAYSTAPGYYSWRNSKDGSWFIQSLCAMLKQYADKLEFMHILTRVNRKVATEFESFSFDATFHAKKQIPCIVSMLTKELYFYH

>casp4

MAEGNHRKKPLKVLESLGKDFLTGVLDNLVEQNVLNWKEEEKKKYYDAKTEDKVRVMADSMQEKQRMAGQMLLQTFFNIDQISPNKKAHPNMEAGPPESGESTDALKLCPHEEFLRLCKERAEEIYPIKERNNRTRLALIICNTEFDHLPPRNGADFDITGMKELLEGLDYSVDVEENLTARDMESALRAFATRPEHKSSDSTFLVLMSHGILEGICGTVHDEKKPDVLLYDTIFQIFNNRNCLSLKDKPKVIIVQACRGANRGELWVRDSPASLEVASSQSSENLEEDAVYKTHVEKDFIAFCSSTPHNVSWRDSTMGSIFITQLITCFQKYSWCCHLEEVFRKVQQSFETPRAKAQMPTIERLSMTRYFYLFPGN

>casp5

MAEDSGKKKRRKNFEAMFKGILQSGLDNFVINHMLKNNVAGQTSIQTLVPNTDQKSTSVKKDNHKKKTVKMLEYLGKDVLHGVFNYLAKHDVLTLKEEEKKKYYDTKIEDKALILVDSLRKNRVAHQMFTQTLLNMDQKITSVKPLLQIEAGPPESAESTNILKLCPREEFLRLCKKNHDEIYPIKKREDRRRLALIICNTKFDHLPARNGAHYDIVGMKRLLQGLGYTVVDEKNLTARDMESVLRAFAARPEHKSSDSTFLVLMSHGILEGICGTAHKKKKPDVLLYDTIFQIFNNRNCLSLKDKPKVIIVQACRGEKHGELWVRDSPASLALISSQSSENLEADSVCKIHEEKDFIAFCSSTPHNVSWRDRTRGSIFITELITCFQKYSCCCHLMEIFRKVQKSFEVPQAKAQMPTIERATLTRDFYLFPGN