UNIVERZITA SV. CYRILA A METODA V TRNAVE

Fakulta prírodných vied

Štefan Janeček

BIOINFORMATIKA PROTEÍNOV

Trnava 2020

Autor: Doc. Ing. Štefan Janeček, DrSc.

Publikácia bola schválená edičnou radou Univerzity sv. Cyrila a Metoda v Trnave a vedením Fakulty prírodných vied Univerzity sv. Cyrila a Metoda v Trnave ako učebné texty pre študentov vysokých škôl.

Vydala: Univerzita sv. Cyrila a Metoda v Trnave, 2020.

Vydanie: prvé.

Web: http://fpv.ucm.sk/images/ucebne_texty/Bioinformatika_proteinov.pdf

ISBN 978-80-572-0085-7

Predslov

Čo je to bioinformatika? Čím sa zaoberá? Čomu sa primárne venuje a čo ponúka? Aký je jej zmysel?

Bioinformatika je vedná disciplína na rozhraní viacerých vedných odborov, ktoré sa vzájomne prelínajú. Ide najmä o prienik matematických odborov a tzv. odborov „Life Sciences“, t.j. vied o živej prírode. Je to teda informatika a aplikovaná matematika na jednej strane, dopĺňané na strane druhej molekulárnou biológiou, genetikou, genomikou, proteomikou, biochémiou, prípadne mikrobiológiou a ďalšími, ako napr. výpočtová biológia, systémová biológia a teoretická biológia. V podstate ide o riešenie biologických problémov na molekulárnej úrovni, ale samotná práca sa vykonáva spôsobom *in silico*, t.j. ani *in vivo* (priamo na živých organizmoch), ani *in vitro* (v laboratóriu; „v skúmavke“), ale v laboratóriu „virtuálnom“ – na počítači.

Pri bioinformatike ide o často o sofistikované získavanie, detailnú analýzu, relevantnú vizualizáciu a správnu interpretáciu zväčša mimoriadne objemných súborov biologických dát, ktoré predstavujú nukleotidové a aminokyselinové sekvencie DNA a RNA, resp. proteínov, ako aj terciárne štruktúry týchto biomakromolekúl, hlavne proteínov. V širšom ponímaní patrí do bioinformatiky aj štúdium údajov o aktivite a expresii génov, pričom významnou súčasťou bioinformatiky je predikcia štruktúr proteínov a ich funkcií, vyúsťujúca vo formulovaní záverov súvisiacich s evolúciou. Výsledky bioinformatiky sú využívané experimenálne orientovaným výskumom, ale aj praxou, napr. v medicíne a polícii. V ideálnom prípade výsledky získané bioinformatickými, t.j. *in silico* prístupmi môžu pomáhať v experimentálnom výskume tým, že naznačujú smer jeho ďalšieho postupu, čím jednoznačne prispevajú k šetreniu finančných zdrojov a skráteniu času potrebného pre dosiahnutie stanovených cieľov.

Ako je zrejmé z názvu, tento učebný text je zameraný na bioinformatické prístupy k štúdiu proteínov.

Obsah

1. Počiatky genómových sekvenačných projektov

Počiatky genómových sekvenačných projektov spadajú do druhej polovice 90. rokov minulého storočia. Išlo, resp. ide o projekty zamerané na získanie nukleotidovej sekvencie kompletného genómu daného organizmu. Zisk tejto informácie je v podstate nevyhnutnou, aj keď nie jedinou a dostačujúcou podmienkou, ktorá umožní komletné porozumenie biológie organizmu.

Sekvenovaniu kompletných genómov organizmov predchádzalo sekvenovanie jednotlivých génov a väčších úsekov DNA, ale potom najmä genómov vírusov a organel. Prvým kompletne sekvenovaným bol bakteriofág φX174 v roku 1977 s 5 386 bázovými pármi (bp). Pre porovnanie: prvý kompletne osekvenovaný, samostatne existujúci organizmus, ktorým bola v roku 1995 baktéria *Haemophilus influeenzae*, mal veľkosť genómu 1 830 837 bp (t.j. ~1,83 Mbp) a 1 743 génov kódujúcich proteíny.

Preto, aby sa mohlo vôbec pristúpiť k začatiu sekvenovania kompletných genómov organizmov, bolo potrebné zhromaždiť nemalé finančné prostriedky a dať dohromady aj obrovký ľudský potenciál. Jedným z najvýznamnejších ľudí, ktorý sa pričinil o spustenie a úspech genómových sekvenačných projektov, bol americký biotechnológ a vizionár Craig Venter, zakladateľ viacerých inštitúcií (napr. The Institute for Genomic Research, TIGR; ale aj iné), v ktorých sa úspešne sekvenovali genómy. Okrem financií a ľudského potenciálu boli súčasťou rozvoja sekvenovania genómov aj pokroky v prístupoch a prístrojovom vybavení v oblasti „Life Sciences“, najmä molekulárnej biológie, v zdokonaľovaní výpočtovej techniky a programových nástrojov, ako aj objavenie a rozmach internetu.

Najskôr sa samozrejme pristúpilo k sekvenovaniu prokaryotov, pretože majú menší jednoduchší genóm v porovnaní s eukaryotmi. Tiež boli medzi prvými sekvenované rôzne modelové a široko využívané mikroorganizmy, ako napr. *Bacillus subtilis* a *Escherichia coli*, ale aj rôzne patogény, ako samotný prvý osekvenovaný genóm *Haemophilus influeenzae* (rôzne infekcie až sepsy), ďalej napr. *Helicobacter pylori* (gastritídy), *Borrelia burgdorferi* (lymská borelióza), *Mycobacterium tuberculosis* (tuberkulóza), *Treponema pallidum* (syfilis), *Vibrio cholerae* (cholera), *Yersinia pestis* (mor) a iné. Ako predstaviteľ

minimalistického genómu bola ešte v roku 1995 osekvenovaná *Mycoplasma genitalium* s veľkosťou genómu 580 070 bp (t.j. 0,58 Mbp) a iba 470 génmi.

Ďalším nemenej významným úspechom genómových sekvenačných projektov bolo definitívne potvrdenie troch domén života v roku 1996 – konkrétne domény *Archaea* – po osekvenovaní prvého archeónu *Methanococcus jannaschii*, ktorý produkuje metán. Archaebaktérie ako samostatnú vývojovú vetvu prokaryotov – odlišnú od „pravých" baktérií – objavil a popísal už v roku 1977 americký mikrobiológ Carl Woese. Ďalším v poradí bol v roku 1997 sírany redukujúci archeón *Archaeoglobus fulgidus*.

Bežný prokaryotický genóm má v hrubom priblížení veľkosť okolo 2 Mbp a obsahuje cca 2 000 génov; samozrejme to je iba veľmi približný, aj keď viac-menej reálny odhad, pre ktorý existujú výnimky nad aj pod ním. Prvou osekvenovanou hubou – eukaryot, ale stále mikroorganizmus – bola v roku 1996 kvasinka *Saccharomyces cerevisiae* s 5 885 génmi a 12,07 Mbp. U eukaryotických organizmov pri narastajúcej veľkosti ich genómov vo vzťahu k menej adekvátnemu nárastu počtu ich génov treba uvažovať ich exón-intrónovú organizáciu s nekódujúcimi oblasťami. V tomto smere je mimoriadne zaujímavý genóm človeka, publikovaný prvý raz v roku 2001, ktorého veľkosť je ~2,91 Gbp, pričom dolný odhad počtu génov, ktoré kódujú proteíny, bol stanovený na hranici 30 000. Pozoruhodné je, že ak by bol počet 30 tisíc génov človeka teoreticky vztiahnutý na prokaryotický genóm (2 Mbp / 2000 génov), jeho veľkosť by mala byť iba asi 30 Mbp.

Ilustrácie prvých kompletných sekvenačných projektov:

Fleischmann RD, … & Venter JC (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269:** 496–512. https://doi.org/101126/science7542800

Fraser CM, … & Venter JC (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403. https://doi.org/10.1126/science.270.5235.397

Bult CJ, … & Venter JC (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1073. https://doi.org/10.1126/science.273.5278.1058

Goffeau A, … & Oliver SG (1996) Life with 6000 genes. *Science* **274:** 546, 563–567. https://doi.org/10.1126/science.274.5287.546

C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282:** 2012–2018. https://doi.org/10.1126/science.282.5396.2012

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408:** 796–815. https://doi.org/10.1038/35048692

Gardner MJ, … & Barrell B (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419:** 498-511. https://doi.org/10.1038/nature01097

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69-87. https://doi.org/10.1038/nature04072

Warren WC, … & Wilson RK (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453:** 175-183. https://doi.org/10.1038/nature06936

Publikácia o sekvenovaní genómu včely je ilustráciou zapojenia množstva autorov z 90 pracovísk celého sveta; nižšie je rozpísaná aj ich vzájoná deľba práce.

Honeybee Genome Sequencing Consortium (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443:** 931-949. https://doi.org/10.1038/nature05260

The Honeybee Genome Sequencing Consortium

Overall project leadership: George M. Weinstock[1,2], Gene E. Robinson[7,9,13,14]

Principal investigators: Richard A. Gibbs[1,2], George M. Weinstock[1,2]

Community coordination: George M. Weinstock (leader)[1,2], Gene E. Robinson (leader)[7,9,13,14], Kim C. Worley (leader)[1,2], Jay D. Evans[4], Ryszard Maleszka[6], Hugh M. Robertson[7,9,13,14], Daniel B. Weaver[16]

Annotation section leaders: Martin Beye[17], Peer Bork[18,19], Christine G. Elsik[20], Jay D. Evans[4], Klaus Hartfelder[25], Greg J. Hunt[27], Hugh M. Robertson[7,9,13,14], Gene E. Robinson[7,9,13,14], Ryszard Maleszka[6], George M. Weinstock[1,2], Kim C. Worley[1,2], Evgeny M. Zdobnov[18,28]

Caste development and reproduction: Klaus Hartfelder (leader)[25], Gro V. Amdam[29], Márcia M. G. Bitondi[26], Anita M. Collins[4], Alexandre S. Cristino[30], Jay D. Evans[4], H. Michael G. Lattorff[31], Carlos H. Lobo[24], Robin F. A. Moritz[31], Francis M. F. Nunes[24], Robert E. Page Jr[29], Zilá L. P. Simões[26], Diana Wheeler[32]

EST sequencing: Piero Carninci (leader)[33], Shiro Fukuda[33], Yoshihide Hayashizaki[33], Chikatoshi Kai[33], Jun Kawai[33], Naoko Sakazume[33], Daisuke Sasaki[33], Michihira Tagami[33]

Brain and behaviour: Ryszard Maleszka (leader)[6], Gro V. Amdam[29], Stefan Albert[34], Geert Baggerman[35], Kyle T. Beggs[37], Guy Bloch[38], Giuseppe Cazzamali[41], Mira Cohen[38], Mark David Drapeau[42], Dorothea Eisenhardt[43], Christine Emore[27], Michael A. Ewing[15], Susan E. Fahrbach[48], Sylvain Forêt[6],

Cornelis J. P. Grimmelikhuijzen[41], Frank Hauser[41], Amanda B. Hummon[15], Greg J. Hunt[27], Jurgen Huybrechts[35], Andrew K. Jones[44], Tatsuhiko Kadowaki[55], Noam Kaplan[40], Robert Kucharski[6], Gérard Leboulle[43], Michal Linial[39,40], J. Troy Littleton[45], Alison R. Mercer[37], Robert E. Page Jr[29], Hugh M. Robertson[7,9,13,14], Gene E. Robinson[7,9,13,14], Timothy A. Richmond[15], Sandra L. Rodriguez-Zas[12], Elad B. Rubin[38], David B. Sattelle[44], David Schlipalius[27], Liliane Schoofs[35], Yair Shemesh[38], Jonathan V. Sweedler[13,15], Rodrigo Velarde[7], Peter Verleyen[35], Evy Vierstraete[35], Michael R. Williamson[41]

Development and metabolism: Martin Beye (leader)[17], Seth A. Ament[13], Susan J. Brown[50], Miguel Corona[7], Peter K. Dearden[36], W. Augustine Dunn[52], Michelle M. Elekonich[53], Christine G. Elsik[20], Sylvain Forêt[6], Tomoko Fujiyuki[54], Irene Gattermeier[17], Tanja Gempe[17], Martin Hasselmann[17], Tatsuhiko Kadowaki[55], Eriko Kage[54], Azusa Kamikouchi[54], Takeo Kubo[54], Robert Kucharski[6], Takekazu Kunieda[54], Marcé Lorenzen[49], Ryszard Maleszka[6], Natalia V. Milshina[20], Mizue Morioka[54], Kazuaki Ohashi[54], Ross Overbeek[57], Robert E. Page Jr[29], Hugh M. Robertson[7,9,13,14], Gene E. Robinson[7,9,13,14], Christian A. Ross[53], Morten Schioett[17], Teresa Shippy[51], Hideaki Takeuchi[54], Amy L. Toth[14], Judith H. Willis[52], Megan J. Wilson[36]

Comparative and evolutionary analysis: Hugh M. Robertson (leader)[7,9,13,14], Evgeny M. Zdobnov (leader)[18,28], Peer Bork[18,19], Christine G. Elsik[20], Karl H. J. Gordon[46], Ivica Letunic[18]

Funding agency management: Kevin Hackett[5], Jane Peterson[58], Adam Felsenfeld[58], Mark Guyer[58]

Physical and genetic mapping: Michel Solignac (leader)[56], Richa Agarwala[59], Jean Marie Cornuet[60], Christine G. Elsik[20], Christine Emore[27], Greg J. Hunt[27], Monique Monnerot[56], Florence Mougel[56], Justin T. Reese[20], David Schlipalius[27], Dominique Vautrin[56], Daniel B. Weaver[16]

Ribosomal RNA genes and related retrotransposable elements: Joseph J. Gillespie (leader)[21,62], Jamie J. Cannone[61], Robin R. Gutell[61], J. Spencer Johnston[21]

Gene prediction and consensus gene set: Christine G. Elsik (leader)[20], Giuseppe Cazzamali[41], Michael B. Eisen[63,64], Cornelis J. P. Grimmelikhuijzen[41], Frank Hauser[41], Amanda B. Hummon[15], Venky N. Iyer[63], Vivek Iyer[65], Peter Kosarev[66], Aaron J. Mackey[67], Ryszard Maleszka[6], Justin T. Reese[20], Timothy A. Richmond[15], Hugh M. Robertson[7,9,13,14], Victor Solovyev[68], Alexandre Souvorov[59], Jonathan V. Sweedler[13,15], George M. Weinstock[1,2], Michael R. Williamson[41], Evgeny M. Zdobnov[18,28]

Honeybee disease and immunity: Jay D. Evans (leader)[4], Katherine A. Aronstein[69], Katarina Bilikova[70], Yan Ping Chen[4], Andrew G. Clark[72], Laura I. Decanini[4], William M. Gelbart[73], Charles Hetru[74], Dan Hultmark[75], Jean-Luc Imler[74], Haobo Jiang[76], Michael Kanost[51], Kiyoshi Kimura[77], Brian P. Lazzaro[71], Dawn L. Lopez[4], Jozef Simuth[70], Graham J. Thompson[78], Zhen Zou[76]

BAC/fosmid library construction and analysis: Pieter De Jong (leader)[79], Erica Sodergren (leader)[1,2], Miklós Csûrös[87], Aleksandar Milosavljevic[1,2], J.

Spencer Johnston[21], Kazutoyo Osoegawa[79], Stephen Richards[1,2], Chung-Li Shu[79], George M. Weinstock[1,2]

G1C content: Christine G. Elsik (leader)[20], Laurent Duret[80], Eran Elhaik[23], Dan Graur[23], Justin T. Reese[20], Hugh M. Robertson[7,9,13,14]

Transposable elements: Hugh M. Robertson (leader)[7,9,13,14], Christine G. Elsik[20]

Gene regulation including miRNA and RNAi: Ryszard Maleszka (leader)[6], Daniel B. Weaver (leader)[16], Gro V. Amdam[29], Juan M. Anzola[20], Kathryn S. Campbell[73], Kevin L. Childs[20], Derek Collinge[46], Madeline A. Crosby[73], C. Michael Dickens[20], Christine G. Elsik[20], Karl H. J. Gordon[46], L. Sian Grametes[73], Christina M. Grozinger[81], Peter L. Jones[9], Mireia Jorda[89], Xu Ling[8], Beverly B. Matthews[73], Jonathan Miller[1,3], Natalia V. Milshina[20], Craig Mizzen[17], Miguel A. Peinado[89], Justin T. Reese[20], Jeffrey G. Reid[3,22], Hugh M. Robertson[7,9,13,14], Gene E. Robinson[7,9,13,14], Susan M. Russo[73], Andrew J. Schroeder[73], Susan E. St Pierre[73], Ying Wang[9], Pinglei Zhou[73]

Superscaffold assembly: Hugh M. Robertson (leader)[7,9,13,14], Richa Agarwala[59], Christine G. Elsik[20], Natalia V. Milshina[20], Justin T. Reese[20], Daniel B. Weaver[16]

Data management: Kim C. Worley (leader)[1,2], Kevin L. Childs[20], C. Michael Dickens[20], Christine G. Elsik[20], William M. Gelbart[73], Huaiyang Jiang[1,2], Paul Kitts[59], Natalia V. Milshina[20], Justin T. Reese[20], Barbara Ruef[59], Susan M. Russo[73], Anand Venkatraman[20], George M. Weinstock[1,2], Lan Zhang[1,2], Pinglei Zhou[69]

Chromosome structure: J. Spencer Johnston (leader)[21], Gildardo Aquino-Perez[21], Jean Marie Cornuet[60], Monique Monnerot[56], Michel Solignac[56], Dominique Vautrin[56]

Population genetics and SNPs: Charles W. Whitfield (leader)[7,13,14], Susanta K. Behura[7], Stewart H. Berlocher[7,14], Andrew G. Clark[72], Richard A. Gibbs[1,2], J. Spencer Johnston[21], Walter S. Sheppard[82], Deborah R. Smith[83], Andrew V. Suarez[7,11], Neil D. Tsutsui[84], Daniel B. Weaver[16], Xuehong Wei[1,2], David Wheeler[1,2]

Genome assembly: George M. Weinstock (leader)[1,2], Kim C. Worley (leader)[1,2], Paul Havlak[1,2], Bingshan Li[1,2], Yue Liu[1,2], Erica Sodergren[1,2], Lan Zhang[1,2]

(A+T)-rich DNA generation: Martin Beye (leader)[17], Martin Hasselmann[17], Angela Jolivet[1,2], Sandra Lee[1,2], Lynne V. Nazareth[1,2], Ling-Ling Pu[1,2], Rachel Thorn[1,2], George M. Weinstock[1,2]

Tiling arrays: Viktor Stolc (leader)[85], Gene E. Robinson (leader)[7,9,13,14], Ryszard Maleszka[6], Thomas Newman[7], Manoj Samanta[85,86], Waraporn A. Tongprasit[85]

Anti-xenobiotic defence mechanisms: Katherine A. Aronstein (leader)[69], Charles Claudianos (leader)[6,46], May R. Berenbaum[7], Sunita Biswas[6,46], Dirk C. de Graaf[47], Rene Feyereisen[90], Reed M. Johnson[7], John G. Oakeshott[46], Hilary Ranson[88], Mary A. Schuler[10]

DNA sequencing: Donna Muzny (leader)[1,2], Richard A. Gibbs (leader)[1,2], George M. Weinstock (leader)[1,2], Joseph Chacko[1,2], Clay Davis[1,2], Huyen Dinh[1,2], Rachel Gill[1,2], Judith Hernandez[1,2], Sandra Hines[1,2], Jennifer

Hume[1,2], LaRonda Jackson[1,2], Christie Kovar[1,2], Lora Lewis[1,2], George Miner[1,2], Margaret Morgan[1,2], Lynne V. Nazareth[1,2], Ngoc Nguyen[1,2], Geoffrey Okwuonu[1,2], Heidi Paul[1,2], Stephen Richards[1,2], Jireh Santibanez[1,2], Glenford Savery[1,2], Erica Sodergren[1,2], Amanda Svatek[1,2], Donna Villasana[1,2], Rita Wright[1,2]

Affiliations for participants: [1]Human Genome Sequencing Center, [2]Department of Molecular and Human Genetics, and [3]Department of Biochemistry, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. [4]Bee Research Laboratory, BARC-E, and [5]National Program Staff, USDA–Agricultural Research Service, Beltsville, Maryland 20705, USA. [6]ARC Special Centre for the Molecular Genetics of Development, Visual Sciences, Research School of Biological Sciences, The Australian National University, Canberra, Australian Capital Territory 0200, Australia. [7]Department of Entomology, [8]Department of Computer Science, [9]Department of Cell and Developmental Biology, [10]Department of Cell and Structural Biology, [11]Department of Animal Biology, [12]Animal Sciences, [13]Neuroscience Program, [14]Program in Ecology and Evolutionary Biology, and [15]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, USA. [16]Bee Power, L.P., 16484 CR 319, Lynn Grove Road, Navasota, Texas 77868, USA. [17]Heinrich-Heine Universitaet Duesseldorf, Institut fuer Genetik, Universitaetsstrasse 1, 40225 Duesseldorf, Germany. [18]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [19]Max Delbrück Center for Molecular Medicine, Robert-Roessle-Strasse 10, 13125 Berlin-Buch, Germany. [20]Department of Animal Science, and [21]Department of Entomology, Texas A&M University, College Station, Texas 77843, USA. [22]Department of Chemistry, and [23]Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204, USA. [24]Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, [25]Departamento de Biologia Celular e Molecular e Bioagentes Patogênicos, Faculdade de Medicina de Ribeirão Preto, and [26]Departamento de Biologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto 14049-900, Brazil. [27]Department of Entomology, Purdue University, West Lafayette, Indiana 47907, USA. [28]Department of Genetic Medicine and Development, University of Geneva Medical School CMU, 1 rue Michel-Servet, 1211 Geneva, Switzerland. [29]School of Life Sciences, Arizona State University, PO Box 874501, Tempe, Arizona 85287-4501, USA. [30]Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brazil. [31]Institut für Zoologie, Molekulare Ökologie, Martin-Luther-Universität Halle-Wittenberg, Hoher Weg 4, D-06099 Halle (Saale), Germany. [32]Department of Entomology, University of Arizona, Tucson, Arizona 85721-0036, USA. [33]Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center, Yokohama 230-0045, Japan. [34]Institut für Medizinische Strahlenkunde und Zellforschung, Versbacher Strasse 5, 97078 Würzburg, Germany. [35]Laboratory of Developmental Physiology, Genomics and Proteomics, K.U. Leuven, Naamsestraat 59 B-3000 Leuven, Belgium. [36]Laboratory for Evolution and Development, Biochemistry Department, and [37]Zoology Department, University of Otago, PO Box 56, Dunedin, New Zealand. [38]Department of Evolution, Systematics, and Ecology, [39]The Sudarsky Center for Computational Biology, and [40]Department of Biological Chemistry, The Alexander Silberman Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 91904, Israel. [41]Center for Functional and Comparative Insect Genomics, Department of Cell Biology and Comparative Zoology, Institute of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark. [42]Department of Biology, New York University, New York, New York 10003, USA. [43]Neurobiology, FB Biology/Chemistry/Pharmacy, Free University Berlin, Koenigin-Luise-Strasse 28/30, 14195 Berlin, Germany. [44]MRC Functional Genetics Unit, Department of Physiology Anatomy and Genetics, Le Gros Clark Building, University of Oxford, South Parks Road, Oxford OX1 3QX, UK. [45]The Picower Institute for Learning and Memory and the Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [46]CSIRO Entomology, GPO Box 1700, Canberra, Australian Capital Territory 2601, Australia. [47]Laboratory of Zoophysiology, University of Ghent, K. L. Ledeganckstraat 35, B-9000 Ghent, Belgium. [48]Department of Biology, Wake Forest University, Winston-Salem, North Carolina 27109, USA. [49]USDA-ARS-GMPRC, 1515 College Avenue, Manhattan, Kansas 66502, USA. [50]Division of Biology, Ackert Hall, [51]Department of Biochemistry, Kansas State

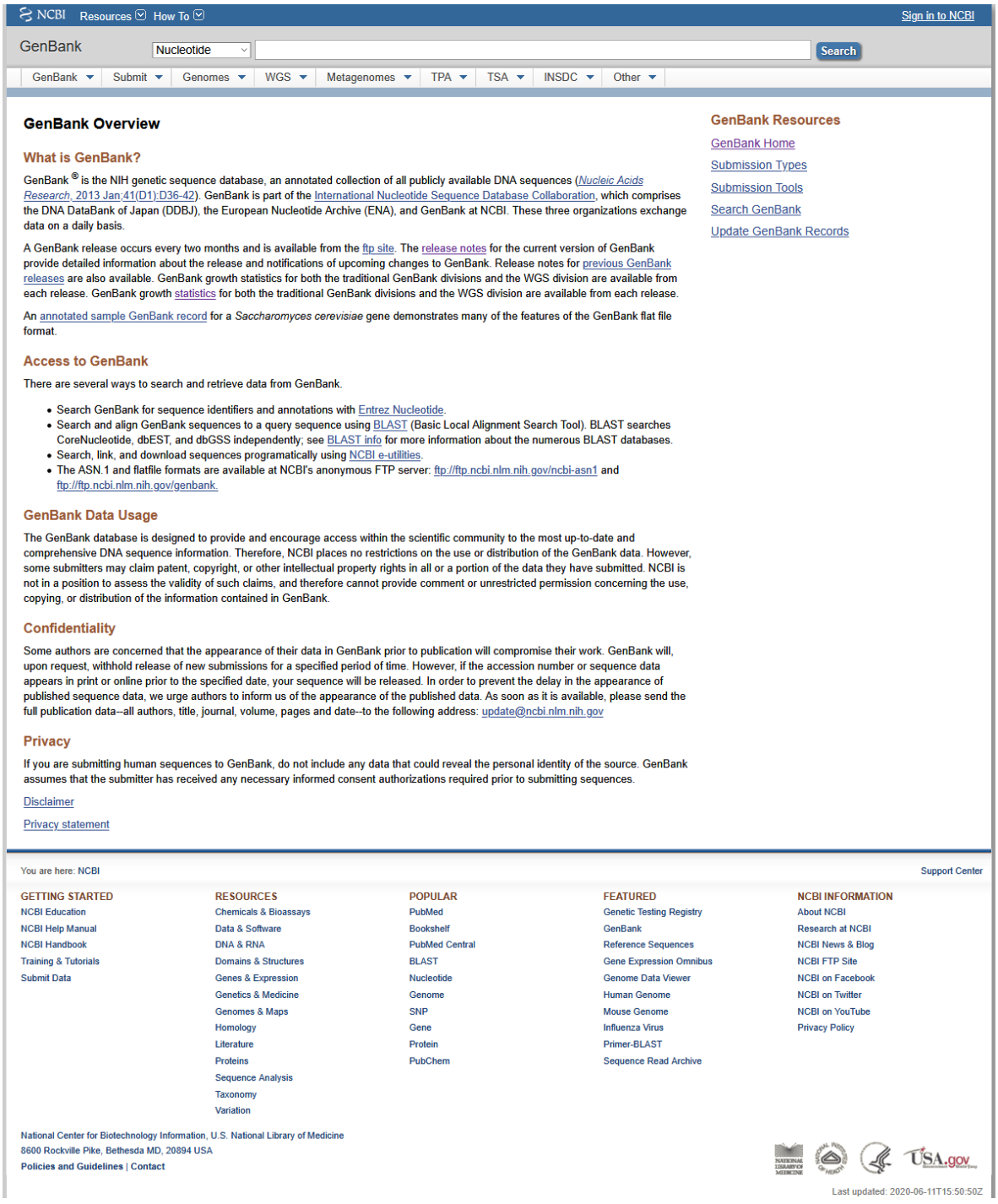
University, Manhattan, Kansas 66506, USA. [52]Department of Cellular Biology, University of Georgia, Athens, Georgia 30602, USA. [53]School of Life Sciences, University of Nevada Las Vegas, 4505 Maryland Parkway, Box 454004, Las Vegas, Nevada 89154-4004, USA. [54]Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan. [55]Graduate School of Bioagricultural Sciences, Nagoya University, Chikusa, Nagoya 464-8601, Japan. [56]Laboratoire Evolution, Génomes et Spéciation Centre National de la Recherche Scientifique, 91198 Gif-sur-Yvette, France. [57]Fellowship for Interpretation of Genomes, 15W155 81st Street, Burr Ridge, Illinois 60527, USA. [58]US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA. [59]National Center for Biotechnology Information, National Library of Medicine, Department of Health and Human Services, 8600 Rockville Pike, Bethesda, Maryland 20894, USA. [60]Centre de Biologie et de Gestion des Populations, Institut National de la Recherche Agronomique, 34988 Saint-Gély-du-Fesc, France. [61]Institute for Cellular and Molecular Biology and Section of Integrative Biology, University of Texas, Austin, Texas 78712, USA. [62]Virginia Bioinformatics Institute 0477, Bioinformatics Facility, Washington Street, Virginia Tech, Blacksburg, Virginia 24061, USA. [63]Division of Genetics, Genomics and Development, Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [64]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA. [65]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [66]Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, New York 10549, USA. [67]Penn Genomics Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [68]Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK. [69]Honey Bee Unit, USDA-ARS, 2413 East highway 83, Number 213, Weslaco, Texas 78596, USA. [70]Slovak Academy of Sciences, Dubravska cesta 21, 845 51 Bratislava 45, Slovakia. [71]Depatment of Entomology, and [72]Department of Molecular Biology and Genetics, Cornell University, Ithaca 14853, New York, USA. [73]Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA. [74]Institut de Biologie Moléculaire et Cellulaire, CNRS, 15 rue René Descartes, 67084 Strasbourg Cedex, France. [75]Umeå Centre for Molecular Pathogenesis, By. 6L, Umeå University, S-901 87 Umeå, Sweden. [76]Department of Entomology and Plant Pathology, Oklahoma State University, 127 NRC, Stillwater, Oklahoma 74078, USA. [77]National Institute of Livestock and Grassland Science, 3-1-1 Kannondai, Tsukuba, Ibaraki, 305-8517, Japan. [78]School of Biological Sciences, University of Sydney, New South Wales 2006, Australia. [79]BACPAC Resources, Children's Hospital Oakland Research Institute, Oakland, California 94609, USA. [80]Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, CNRS, Univ. Lyon 1, 69622 Villeurbanne Cedex, France. [81]Department of Entomology, W.M. Keck Center for Behavioral Biology, Gardner Hall, MC 7613, North Carolina State University, Raleigh, North Carolina 27695, USA. [82]Department of Entomology, Washington State University, Pullman, Washington 99164, USA. [83]Department of Ecology & Evolutionary Biology/Entomology, Haworth Hall, 1200 Sunnyside Avenue, University of Kansas, Lawrence, Kansas 66045, USA. [84]Department of Ecology and Evolutionary Biology, University of California, Irvine, 321 Steinhaus Hall, Irvine, California 92697, USA. [85]NASA Ames Genome Research Facility, Moffet Field, California 94035, USA. [86]Systemix Institute, Cupertino, California 95014, USA. [87]Departement d'informatique et de recherche operationnelle, Universite de Montreal, CP 6128 succ. Centre-Ville, Montreal, Quebec H3C 3J7, Canada. [88]Vector Research, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK. [89]Research Institute of Oncology, L'Hospitalet 08907, Catalonia, Spain. [90]Institut National de la Recherche Agronomique and Université de Nice Sophia Antipolis, UMR 1112, Centre de Recherche de Sophia Antipolis, 06903 Sophia Antipolis, France.

2. GenBank a UniProt – základné sekvenčné databázy

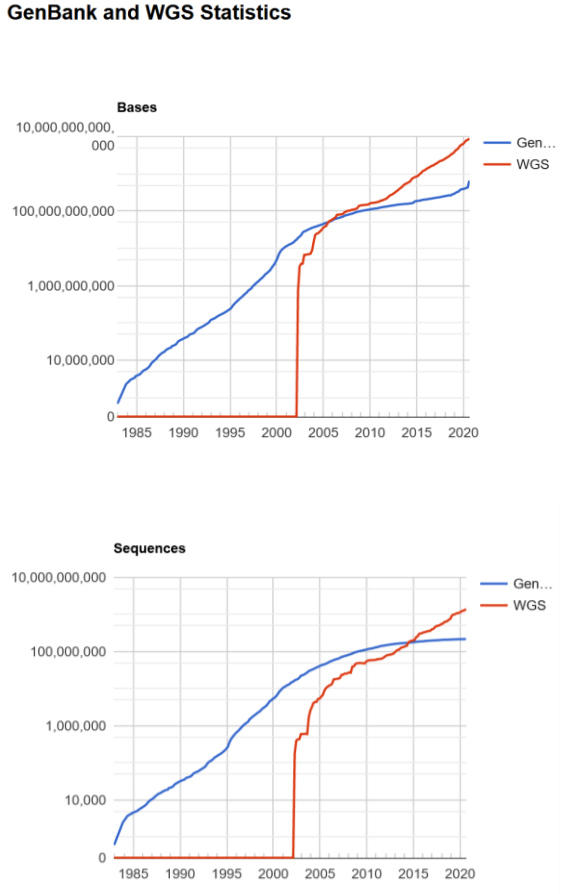
GenBank NCBI

Úvodná web-stránka databázy GenBank:

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



Databáza GenBank je genetická sekvenčná databáza v rámci NIH (National Institutes of Health) na tzv. NCBI serveri (National Center for Biotechnology Information). Je to anotovaná kolekcia všetkých verejne dostupných DNA sekvencií. V aktuálnom vydaní – august 2020 – GenBank obsahuje viac ako 650 miliárd báz v ~218 miliónoch sekvenčných záznamoch (Obr. 2.1).



GENBANK AND WGS STATISTICS

		GenBank		WGS	
Release	Date	Bases	Sequences	Bases	Sequences
3	Dec 1982	680338	606		
14	Nov 1983	2274029	2427		
20	May 1984	3002088	3665		
24	Sep 1984	3323270	4135		
25	Oct 1984	3368765	4175		
26	Nov 1984	3689752	4393		
32	May 1985	4211931	4954		
36	Sep 1985	5204420	5700		
40	Feb 1986	5925429	6642		
42	May 1986	6765476	7416		
44	Aug 1986	8442357	8823		
46	Nov 1986	9615371	9978		
48	Feb 1987	10961380	10913		
50	May 1987	13048473	12534		
52	Aug 1987	14855145	14020		
53	Sep 1987	15514776	14584		
54	Dec 1987	16752872	15465		
.....					
224	Feb 2018	253630708098	207040555	2608532210351	564286852
225	Apr 2018	260189141631	208452303	2784740996536	621379029
226	Jun 2018	263957884539	209775348	2944617324086	639804105
227	Aug 2018	260806936411	208831050	3204855013281	665309765
228	Oct 2018	279668290132	209656636	3444172142207	722438528
229	Dec 2018	285688542186	211281415	3656719423096	773773190
230	Feb 2019	303709510632	212260377	4164513961679	945019312
231	Apr 2019	321680566570	212775414	4421986382065	993732214
232	Jun 2019	329835282370	213383758	4847677297950	1022913321
233	Aug 2019	366733917629	213865349	5585922333160	1075272215
234	Oct 2019	386197018538	216763706	5985250251028	1097629174
235	Dec 2019	388417258009	215333020	6277551200690	1127023870
236	Feb 2020	399376854872	216214215	6968991265752	1206720688
237	Apr 2020	415770027949	216531829	7788133221338	1267547429
238	Jun 2020	427823258901	217122233	8114046262158	1302852615
239	Aug 2020	654057069549	218642238	8841649410652	1408122887

Obr. 2.1. Štatistika databázy GenBank. Na grafoch vľavo je naznačený nárast dát – báz (hore) a sekvencií (dolu) od roku 1983 priamo v databáze GenBank (modrá farba), resp. ako prírastky z projektov celogenómového sekvenovania – WGS (whole genome shotgun).
V tabuľke vpravo sú údaje o počte báz a sekvencií v databáze GenBank od roku 1982 (hore) a tie isté údaje pre databázu GenBank a prírastky z projektov celogenómového sekvenovania – WGS (whole genome shotgun) – od roku 2018 doteraz (dole).
Zdroj: https://www.ncbi.nlm.nih.gov/genbank/statistics/

GenBank je súčasťou konzorcia *INSDC* (International Nucleotide Sequence Database Collaboration), t.j. medzinárodnej spolupráce nukleotidových sekvenčných databáz:

(i) *GenBank* (v rámci NCBI – National Center for Biotechnology Information, Bethesda, MD, USA);

(ii) *ENA* – European Nucleotide Archive – pôvodne EMBL Nucleotide Database (v rámci EBI – European Bioinformatics Institute, Hinxton, UK);

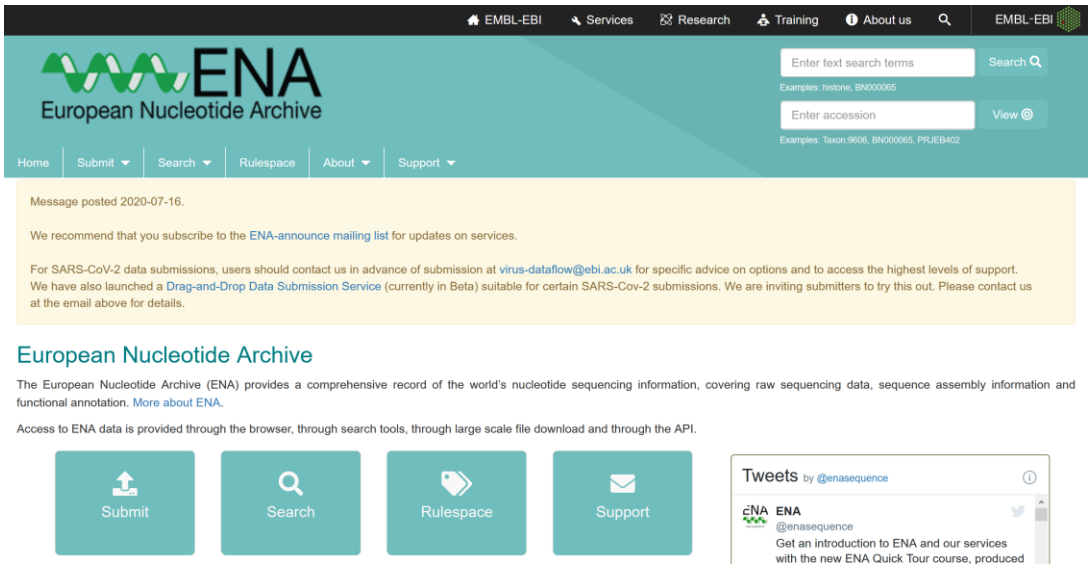
(iii) *DDBJ* – DNA DataBank of Japan (v rámci NIG – National Institute of Genetics, Shizuoka, Japan).

Úvodná web-stránka konzorcia INSDC:
http://www.insdc.org/

Sekvencia v databáze je jednoznačne charakterizovaná tzv. prístupovým číslom (Accession No.). Prístupové čísla pre tú istú sekvenciu sú vo všetkých troch nukleotidových databázach – GenBank, ENA a DDBJ – rovnaké, pretože tieto tri databázy sú obsahovo totožné. Rozdiel je len v ich vizuálnom spracovaní a v tom, že ich spravujú rôzni kurátori v rámci rôznych inštitúcií. Dáta v týchto troch databázach sú rovnaké. V podstate je jedno, do ktorej databázy sú nové dáta vložené najskôr, keďže sa spolupráca, t.j. aj vzájomná výmena dát, uskutočňuje na dennej báze.

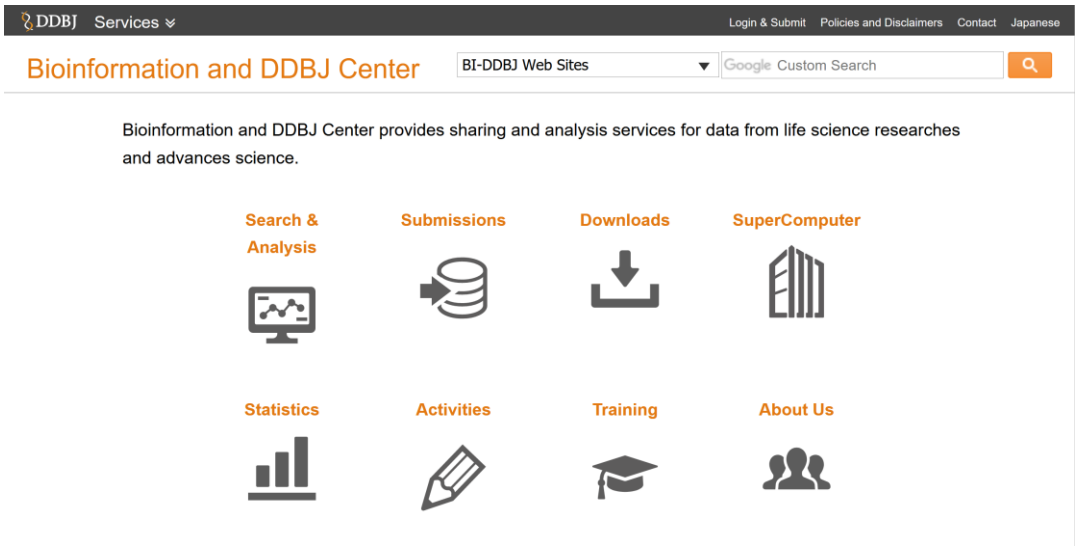
Úvodná web-stránka databázy ENA:

https://www.ebi.ac.uk/ena/



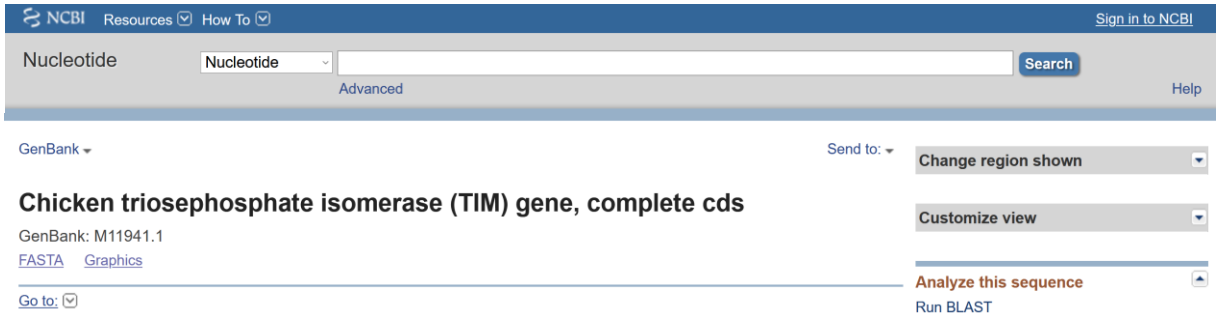
Úvodná web-stránka databázy DDBJ – DNA Data Bank of Japan:

https://www.ddbj.nig.ac.jp/



V nukleotidových databázach sa okrem samotnej nukleotidovej sekvencie nachádza aj jej preklad („translation“), t.j. aminokyselinová sekvencia. V dôsledku toho, že jeden najmä dlhší nukleotidový úsek môže kódovať viac proteínov, jednoznačnejším indikátorom danej sekvencie je tzv. „protein_id“, ktoré predstavuje prístupové číslo konkrétnej aminokyselinovej sekvencie v pod-databáze GenPept v rámci celej databázy GenBank.

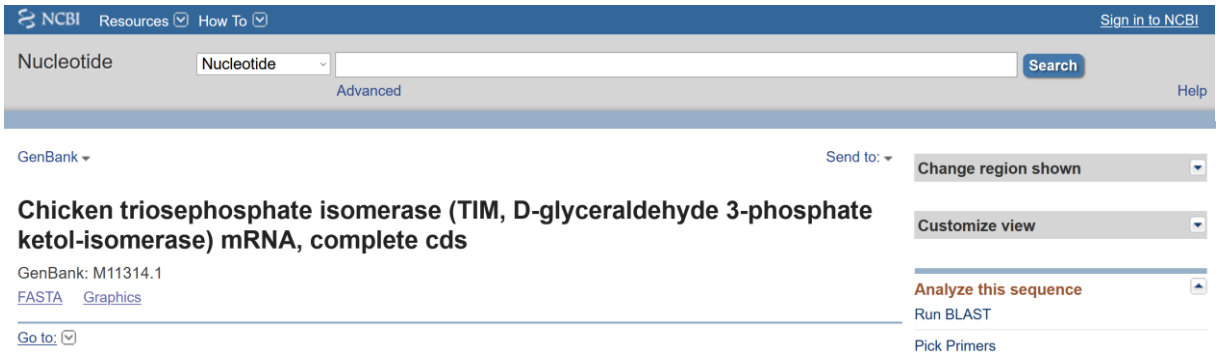
Záznam v databáze GenBank pre DNA kuracej triózafosfátizomerázy (prístupové číslo: M11941; protein_id: AAA49095.1):




```
LOCUS       CHKTIMA                 3900 bp    DNA     linear   VRT 28-APR-1993
DEFINITION  Chicken triosephosphate isomerase (TIM) gene, complete cds.
ACCESSION   M11941
VERSION     M11941.1
KEYWORDS    D-glyceraldehyde 3-phosphate ketol-isomerase; triose-phosphate
            isomerase.
SOURCE      Gallus gallus (chicken)
  ORGANISM  Gallus gallus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
            Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes;
            Phasianidae; Phasianinae; Gallus.
REFERENCE   1  (bases 1 to 3900)
  AUTHORS   Straus,D. and Gilbert,W.
  TITLE     Genetic engineering in the Precambrian: structure of the chicken
            triosephosphate isomerase gene
  JOURNAL   Mol Cell Biol 5 (12), 3497-3506 (1985)
   PUBMED   3837846
COMMENT     Original source text: Chicken embryonic muscle DNA.
FEATURES             Location/Qualifiers
     source          1..3900
                     /organism="Gallus gallus"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9031"
     prim_transcript 273..3198
                     /note="TIM mRNA and introns"
     CDS             join(324..435,1269..1392,1487..1571,1707..1839,2070..2155,
                     2292..2379,2634..2752)
                     /note="triosephosphate isomerase (EC 5.3.1.1)"
                     /codon_start=1
                     /protein_id="AAA49095.1"
                     /translation="MAPRKFFVGGNWKMNGDKKSLGELIHTLNGAKLSADTEVVCGAP
                     SIYLDFARQKLDAKIGVAAQNCYKVPKGAFTGEISPAMIKDIGAAWVILGHSERRHVF
                     GESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKAIADNVKDWSKV
                     VLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSHVSDAVAQSTRIIYGGSVTGGNCK
                     ELASQHDVDGFLVGGASLKPEFVDIINAKH"
     exon            <324..435
                     /note="triosephosphate isomerase (EC 5.3.1.1)"
                     /number=1
     intron          436..1268
                     /note="TIM, intron A"
     exon            1269..1392
                     /note="triosephosphate isomerase"
                     /number=2
     intron          1393..1486
                     /note="TIM, intron B"
     exon            1487..1571
                     /note="triosephosphate isomerase"
                     /number=3
     intron          1572..1706
                     /note="TIM, intron C"
     exon            1707..1839
                     /note="triosephosphate isomerase"
                     /number=4
     intron          1840..2069
                     /note="TIM, intron D"
     exon            2070..2155
                     /note="triosephosphate isomerase"
```

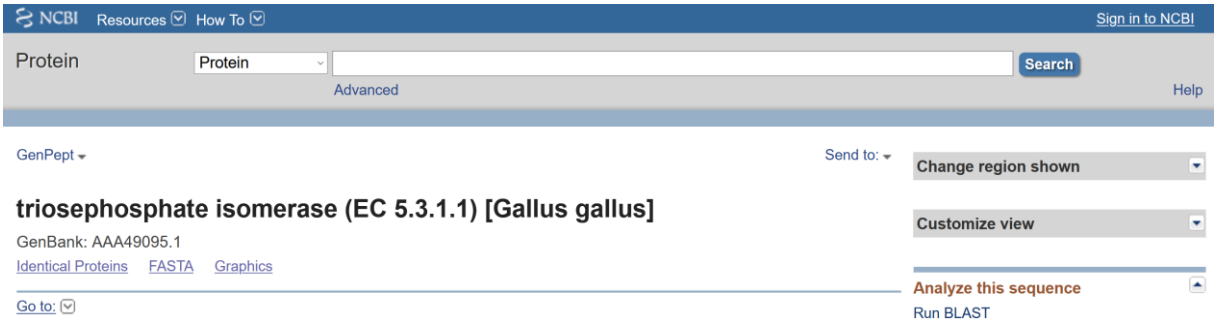
```
     intron          2156..2291
                     /note="TIM, intron E"
     exon            2292..2379
                     /note="triosephosphate isomerase"
                     /number=6
     intron          2380..2633
                     /note="TIM, intron F"
     exon            2634..>2752
                     /note="triosephosphate isomerase"
                     /number=7
ORIGIN
        1 ggggccccgc gagccgaaac tcgggcactg ccgcggtacg agcttccgta cggagaaggg
       61 aggtcgcgtg gttctcgtgt gggggcgcac ggagcttcgg tgcgggcttc gcgcccagtg
      121 cagcttctga ggaagggccg cggccatctc cgagcggagg ccgcccggct gccgaggccc
      181 agggggtggg gagcgagccg gggggcgggg ctacgggggg tggggcttcg cgccccgccg
      241 gcctataaaa gcggccgccg cggctccgtg ccgttgccga ccttcgcctg cgccgctgct
      301 gcttcgcgcc cgtcgcctcc gccatggctc ccaggaagtt cttcgtgggt ggcaactgga
      361 agatgaacgg cgacaagaag agcttgggcg agctcatcca cacgctgaat ggcgccaagc
      421 tctcggccga caccggtgag ggagggcgcg gtcgcagccg cacgtagcgc acgtcccgcg
      481 cggcggggcc cggcccggca ggggcacggg aaatgtcagg gggcggcggg gccgaaactt
      541 aaccccgcgg cggggctaaa aataccgcga caccggatcc ctgccgctga agggccccgc
      601 atcccaccct cccctccccc cgccgccgcc ggatgcgggc agcccccggc cgccggcgcc
      661 gcgggggtct gcgggcccgg cccgcgagga gactgcggcg ggcggtgagg ggcctccggg
      721 ctccgtccct ccgcgcggag ctgctgtccc cgcggccccg tctctacggc cccggcctgc
      781 cgcccaacct cgcgcccggg gtcgggggcg cgatatttgg gccgaggtca cggcgctgcg
      841 gagcgggttc caggagcgtg cctagcctgc cgcggtgccg tacccgggtg taatccgcgt
      901 gtaacccgac gctgcccttt tctcccccgt tcagccttgc ggccgggtga gattcttgcc
      961 cggtttctcg gctgggatgt ttccaaaggg ctcggtgtga actggctgga ggcctggctc
     1021 tctgcctgac tccccccccgc tgacccctgg agtgagtttc ccctccgctg tggagtcatt
     1081 ggatagtagg gaggatgttc ctactctgtg ggtcccccaa atgtcagcga gggcaggctg
     1141 agcctccacc aagtgggggc tgggtggggg agggggggagg ggtggcctca gggctgggcc
     1201 aaggggccga aggttgcgtg cagcttctat cccttggcct tctgatggat gacctttctg
     1261 ccttacagag gtggtttgcg gagccccttc aatctacctt gattttgccc gccagaagct
     1321 tgatgcaaag attggagttg cagcacaaaa ctgttacaag gtaccgaagg gtgctttcac
     1381 aggagagatc aggtgagccc aaagtgaaga ttaaagcaag cattgctttg cagagcaaac
     1441 ttgctctgga gatgtgagtc actaccttca tttgctttct catcagccca gcaatgatca
     1501 aagatattgg agctgcatgg gtgatcctgg gccactcaga gcggaggcat gtttttggag
     1561 agtctgatga ggtgagagac tctggagtgg agaggtggta ccaaacagat aattctgtga
     1621 aggcaggagg agtgagcggt ttgtgggaaa ggtgggtaca ggagcaggtc tgactgccac
     1681 ttaccttatt cctcttccat ccacagttga ttgggcagaa ggtggctcat gctcttgctg
     1741 aaggcctcgg tgtcatcgcc tgcattgggg agaagctgga tgagagagaa gctggcataa
     1801 cggagaaggt ggtctttgaa cagaccaaag ctattgctgg taagagaaga aaactgactt
     1861 cttttctttg gacagctggg aaatggctat tgtggccata actgcctgtt cctggccttc
     1921 cttgagggct ttcagataca ttgctacaaa tgtttacaaa cagtggtcgc ctttgtagaa
     1981 ggtctgcatg ttccaggagg ctttttttcca gatctgccat tcccaacatg aggtgctgta
     2041 ctcatgtata ccttcttctt tcccctcaga taacgtgaag gactggagta aggtggttct
     2101 tgcctatgag ccagtttggg ctatcggaac tggtaaaact gctactcccc aacaggtatt
     2161 agcaagggaa aggtggctga tgaaatgact gtctgacccc actcaagagg tctccctgaa
     2221 gatggaacaa ggcagttcct tctttggaag ggaattgtat gcaggatggt tcagattcta
     2281 tttcttttca ggctcaggag gttcatgaga agctgagagg ctggctcaaa agccacgtgt
     2341 ctgatgctgt tgctcagtca actaggatca tctatggagg taaatgagtg tatagtctct
     2401 ttgaggctgc acttactttc tttgcttgct gctgttgtac agcatgctct gctcaagagg
     2461 aaacatgaga ctaacagagc atgaagaacc acttactagc ttgtttaggt aaatgggaga
     2521 cagtcttacg attgatagaa taagaggtgg ttagtacctt gcaagtttta cctctgtgtt
     2581 attgccattt tgggtttgcc atgtctgaac tcttaatcct gttccctctt taggttcagt
     2641 cactggtggc aactgtaagg aactggcctc ccagcatgat gtggatggct tccttgttgg
     2701 tggggcttct ctcaagccag agtttgtgga tattatcaat gcaaaacatt aaagcagcct
     2761 gtgaggagca gtcccttacg gttaagagca agaaactgaa gcaagaaggg accttgtgtt
     2821 gcacgtctct cggtacagag gcttcttctg aggctttccc ccaccaccac aattattgtt
     2881 ctagctgtgc tgctaacccc caccaccttg ttggagtccc attagtgtga gcccatctca
     2941 gcagagtctc ctttctgaac tggcaaaatc cttggttatc tgttgagctg ttagagccca
     3001 cagtcaacct ggccattgcc tctctctctt tgcagccctg cagggaggga gggccactag
     3061 tactgggggg aagaaaaagg aaccaccatc ttctgcatct ttcagctcca tccgcaagga
     3121 gcctggcagc ttagaccctt gtgagacacc ttacctcacc aatgtcctgt attgaacaat
     3181 aaatgagaag gaaaaaaagt gtgtcctggt tattatttag aagcaattaa ggacaaggtg
     3241 tatacagata gactggagta ttttggcagt agctagctgc agagggagca gggtgttcaa
     3301 gagttctgtc tttataacct ttccctgggg gtcaatgtct tttttggttg ttacgttttt
     3361 aattcttaca cccctccact tccttgttgc tccaacaagt gttgcacagg tgcaacatgt
     3421 atcctctcac tatacttctc acctagaaca ttcccacaga acctttattt accgttgcag
     3481 ttcctcttgg ctgtatgagg attaacacgc tgcagtgcct tccagtagct gcaagggctg
     3541 tcccactaat gagtgactct tcctgccact gacaccacac tggtggctgc tggctaggtg
     3601 agcctgcaca agtgtgagca atggctgcag gctggtgtga ggccagatgg tgaaactgct
     3661 gctgtggggt agtaataatc ctaaattaat tggcgaggct gcctggccag agattcccat
     3721 caagtaggaa agttggcctg aggaggaatg gagggcgtgc tggtgaagat agagcacgca
     3781 gtggctaggg gtgtgcagaa aagatgtttc aaaagtttaa acggttggag gaaaaagatc
     3841 ctaagatctg tggcttctgc aagcaaagtg ctcaaaaggc ttctggggcg tggctttagt
```

Záznam v databáze GenBank pre mRNA kuracej triózafosfátizomerázy (prístupové číslo: M11314; protein_id: AAA49094.1):



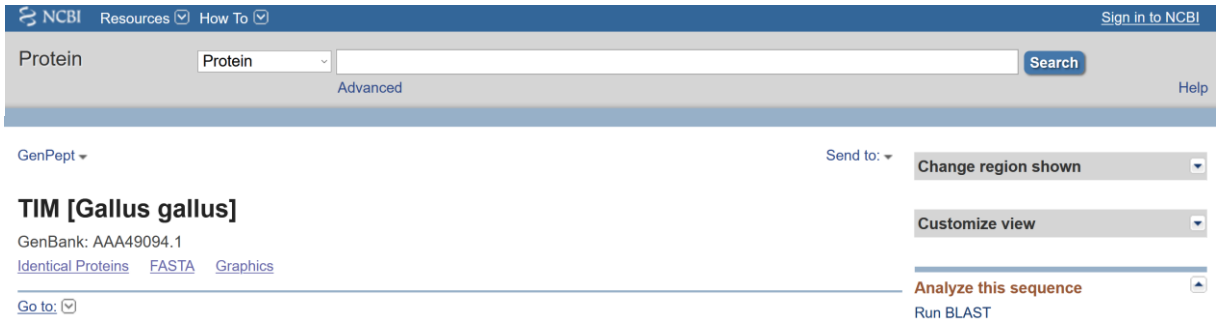

```
LOCUS       CHKTIM                  1303 bp    mRNA    linear   VRT 28-APR-1993
DEFINITION  Chicken triosephosphate isomerase (TIM, D-glyceraldehyde
            3-phosphate ketol-isomerase) mRNA, complete cds.
ACCESSION   M11314
VERSION     M11314.1
KEYWORDS    D-glyceraldehyde 3-phosphate ketol-isomerase; isomerase;
            triose-phosphate isomerase.
SOURCE      Gallus gallus (chicken)
  ORGANISM  Gallus gallus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
            Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes;
            Phasianidae; Phasianinae; Gallus.
REFERENCE   1  (bases 1 to 1303)
  AUTHORS   Straus,D. and Gilbert,W.
  TITLE     Chicken triosephosphate isomerase complements an Escherichia coli
            deficiency
  JOURNAL   Proc Natl Acad Sci U S A 82 (7), 2014-2018 (1985)
   PUBMED   3885220
COMMENT     Original source text: Chicken breast muscle cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..1303
                     /organism="Gallus gallus"
                     /mol_type="mRNA"
                     /db_xref="taxon:9031"
     CDS             118..864
                     /note="TIM"
                     /codon_start=1
                     /protein_id="AAA49094.1"
                     /translation="MAPRKFFVGGNWKMNGDKKSLGELIHTLNGAKLSADTEVVCGAP
                     SIYLDFARQKLDAKIGVAAQNCYKVPKGAFTGEISPAMIKDIGAAWVILGHSERRHVF
                     GESDELIGQKVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKAIADNVKDWSKV
                     VLAYEPVWAIGTGKTATPQQAQEVHEKLRGWLKSHVSDAVAQSTRIIYGGSVTGGNCK
                     ELASQHDVDGFLVGGASLKPEFVDIINAKH"
ORIGIN
        1 ggggctacgg ggggtggggc ttcgcgcccc gccggcctat aaaagcggcc gccgcggctc
       61 cgtgccgttg ccgaccttcg cctgcgccgc tgctgcttcg cgcccgtcgc ctccgccatg
      121 gctcccagga agttcttcgt gggtggcaac tggaagatga acggcgacaa gaagagcttg
      181 ggcgagctca tccacacgct gaatggcgcc aagctctcgg ccgacaccga ggtggtttgc
      241 ggagcccctt caatctacct tgattttgcc cgccagaagc ttgatgcaaa gattggagtt
      301 gcagcacaaa actgttacaa ggtaccgaag ggtgctttca caggagagat cagcccagca
      361 atgatcaaag atattggagc tgcatgggtg atcctgggcc actcagagcg gaggcatgtt
      421 tttggagagt ctgatgagtt gattgggcag aaggtggctc atgctcttgc tgaaggcctc
      481 ggtgtcatcg cctgcattgg ggagaagctg gatgagagag aagctggcat aacggagaag
      541 gtggtctttg aacagaccaa agctattgct gataacgtga aggactggag taaggtggtt
      601 cttgcctatg agccagtttg ggctatcgga actggtaaaa ctgctactcc ccaacaggct
      661 caggaggttc atgagaagct gagaggctgg ctcaaaagcc acgtgtctga tgctgttgct
      721 cagtcaacta ggatcatcta tggaggttca gtcactggtg gcaactgtaa ggaactggcc
      781 tcccagcatg atgtggatgg cttccttgtt ggtggggctt ctctcaagcc agagtttgtg
      841 gatattatca atgcaaaaca ttaaagcagc ctgtgaggag cagtccctta cggttaagag
      901 caagaaactg aagcaagaag ggaccttgtg ttgcacgtct ctcggtacag aggcttcttc
      961 tgaggctttc ccccaccacc acaattattg ttctagctgt gctgctaacc cccaccacct
     1021 tgttggagtc ccattagtgt gagcccatct cagcagagtc tcctttctga actggcaaaa
     1081 tccttggtta tctgttgagc tgttagagcc cacagtcaac ctggccattg cctctctctc
     1141 tttgcagccc tgcagggagg gagggccact agtactgggg ggaagaaaaa ggaaccacca
     1201 tcttctgcat ctttcagctc catccgcaag gagcctggca gcttagaccc ttgtgagaca
     1261 ccttacctca ccaatgtcct gtattgaaca ataaatgaga agg
```

Záznam v databáze GenBank/GenPept pre aminokyselinovú sekvenciu kuracej triózafosfátizomerázy (prístupové číslo: AAA49095.1) – preklad z DNA (prístupové číslo: M11941):




```
LOCUS       AAA49095                 248 aa            linear   VRT 28-APR-1993
DEFINITION  triosephosphate isomerase (EC 5.3.1.1) [Gallus gallus].
ACCESSION   AAA49095
VERSION     AAA49095.1
DBSOURCE    locus CHKTIMA accession M11941.1
KEYWORDS    .
SOURCE      Gallus gallus (chicken)
  ORGANISM  Gallus gallus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
            Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes;
            Phasianidae; Phasianinae; Gallus.
REFERENCE   1  (residues 1 to 248)
  AUTHORS   Straus,D. and Gilbert,W.
  TITLE     Genetic engineering in the Precambrian: structure of the chicken
            triosephosphate isomerase gene
  JOURNAL   Mol Cell Biol 5 (12), 3497-3506 (1985)
   PUBMED   3837846
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..248
                     /organism="Gallus gallus"
                     /db_xref="taxon:9031"
     Protein         1..248
                     /name="triosephosphate isomerase (EC 5.3.1.1)"
     Region          1..246
                     /region_name="PTZ00333"
                     /note="triosephosphate isomerase; Provisional"
                     /db_xref="CDD:240365"
     Site            order(11,13,95,165,171,211,230,232..233)
                     /site_type="other"
                     /note="substrate binding site [chemical binding]"
                     /db_xref="CDD:238190"
     Site            order(11,14,45..47,49,52,64,82,85..86,97..98)
                     /site_type="other"
                     /note="dimer interface [polypeptide binding]"
                     /db_xref="CDD:238190"
     Site            order(13,95,165)
                     /site_type="active"
                     /note="catalytic triad [active]"
                     /db_xref="CDD:238190"
     CDS             1..248
                     /coded_by="join(M11941.1:324..435,M11941.1:1269..1392,
                     M11941.1:1487..1571,M11941.1:1707..1839,
                     M11941.1:2070..2155,M11941.1:2292..2379,
                     M11941.1:2634..2752)"
ORIGIN
        1 maprkffvgg nwkmngdkks lgelihtlng aklsadtevv cgapsiyldf arqkldakig
       61 vaaqncykvp kgaftgeisp amikdigaaw vilghserrh vfgesdelig qkvahalaeg
      121 lgviacigek ldereagite kvvfeqtkai adnvkdwskv vlayepvwai gtgktatpqq
      181 aqevheklrg wlkshvsdav aqstriiygg svtggnckel asqhdvdgfl vggaslkpef
      241 vdiinakh
```

Záznam v databáze GenBank/GenPept pre aminokyselinovú sekvenciu kuracej triózafosfátizomerázy (prístupové číslo: AAA49094.1) – preklad z mRNA (prístupové číslo: M11314):

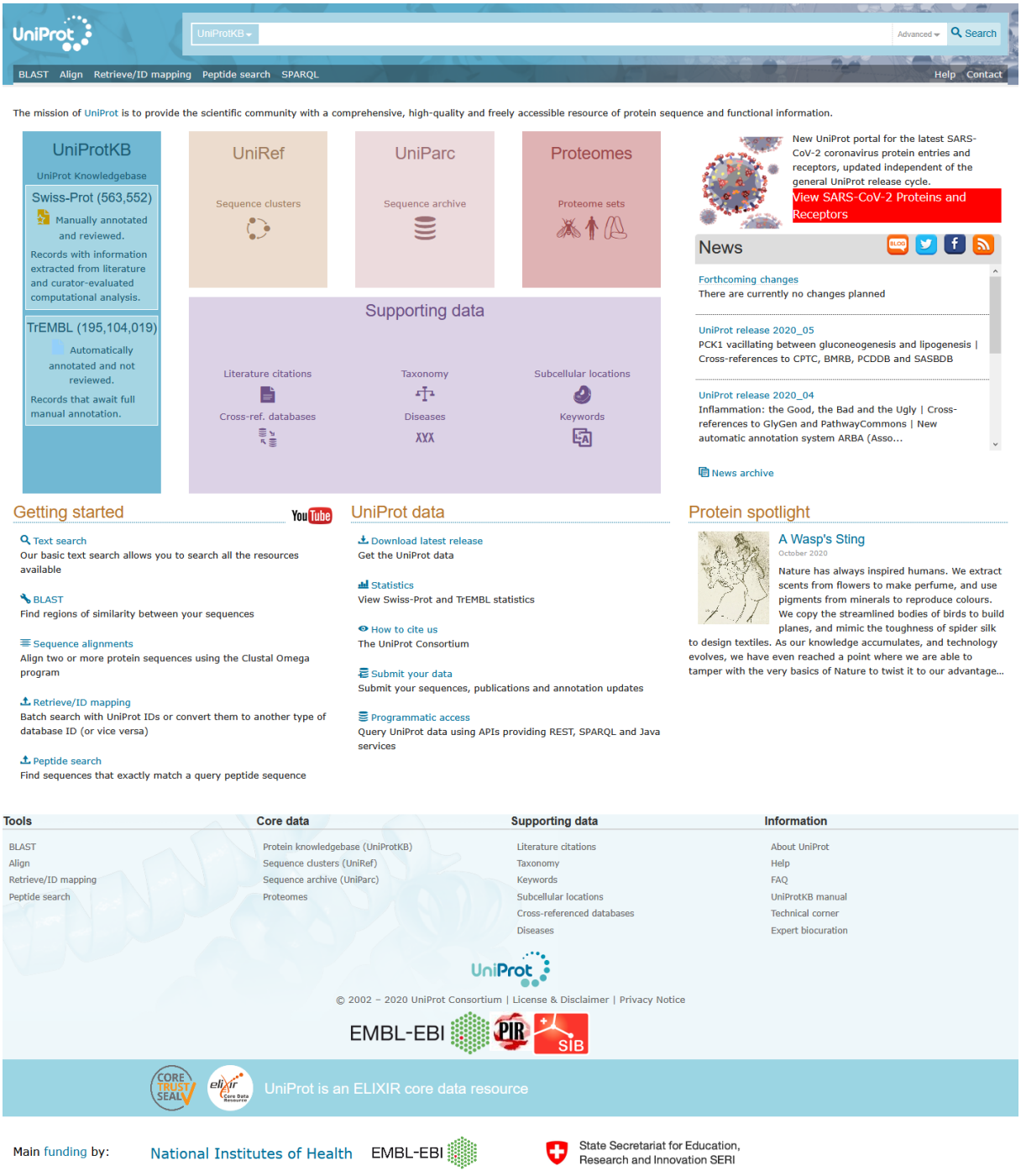



```
LOCUS       AAA49094                 248 aa            linear   VRT 28-APR-1993
DEFINITION  TIM [Gallus gallus].
ACCESSION   AAA49094
VERSION     AAA49094.1
DBSOURCE    locus CHKTIM accession M11314.1
KEYWORDS    .
SOURCE      Gallus gallus (chicken)
  ORGANISM  Gallus gallus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
            Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes;
            Phasianidae; Phasianinae; Gallus.
REFERENCE   1  (residues 1 to 248)
  AUTHORS   Straus,D. and Gilbert,W.
  TITLE     Chicken triosephosphate isomerase complements an Escherichia coli
            deficiency
  JOURNAL   Proc Natl Acad Sci U S A 82 (7), 2014-2018 (1985)
   PUBMED   3885220
COMMENT     Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..248
                     /organism="Gallus gallus"
                     /db_xref="taxon:9031"
     Protein         1..248
                     /name="TIM"
     Region          1..246
                     /region_name="PTZ00333"
                     /note="triosephosphate isomerase; Provisional"
                     /db_xref="CDD:240365"
     Site            order(11,13,95,165,171,211,230,232..233)
                     /site_type="other"
                     /note="substrate binding site [chemical binding]"
                     /db_xref="CDD:238190"
     Site            order(11,14,45..47,49,52,64,82,85..86,97..98)
                     /site_type="other"
                     /note="dimer interface [polypeptide binding]"
                     /db_xref="CDD:238190"
     Site            order(13,95,165)
                     /site_type="active"
                     /note="catalytic triad [active]"
                     /db_xref="CDD:238190"
     CDS             1..248
                     /coded_by="M11314.1:118..864"
ORIGIN
        1 maprkffvgg nwkmngdkks lgelihtlng aklsadtevv cgapsiyldf arqkldakig
       61 vaaqncykvp kgaftgeisp amikdigaaw vilghserrh vfgesdelig qkvahalaeg
      121 lgviacigek ldereagite kvvfeqtkai adnvkdwskv vlayepvwai gtgktatpqq
      181 aqevheklrg wlkshvsdav aqstriiygg svtggnckel asqhdvdgfl vggaslkpef
      241 vdiinakh
```

UniProt

Úvodná web-stránka databázy UniProt:

http://www.uniprot.org/

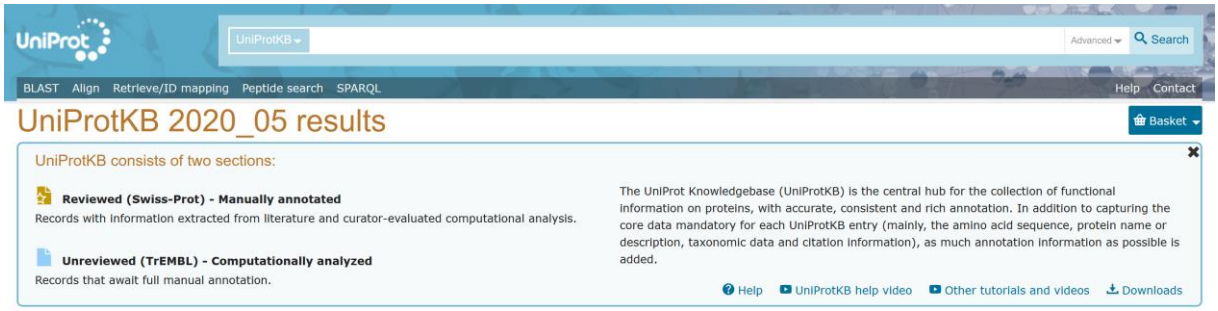


Databáza UniProt KnowledgeBase (UniProt KB) je centrom pre kolekciu informácií o proteínoch so správnou, konzistentnou a bohatou anotáciou o ich funkciách. Je spravovaná v rámci EBI (European Bioinformatics Institute), ktorý predstavuje „vlajkovú loď“ v oblasti bioinformatiky v Európe.

UniProtKB pozostáva z dvoch základných súčastí:

(i) SwissProt – manuálne anotované záznamy (recenzované záznamy);

(ii) TrEMBL – automaticky anotované, t.j. počítačovo analyzované záznamy (nerecenzované záznamy, očakávajúce plnú anotáciu).

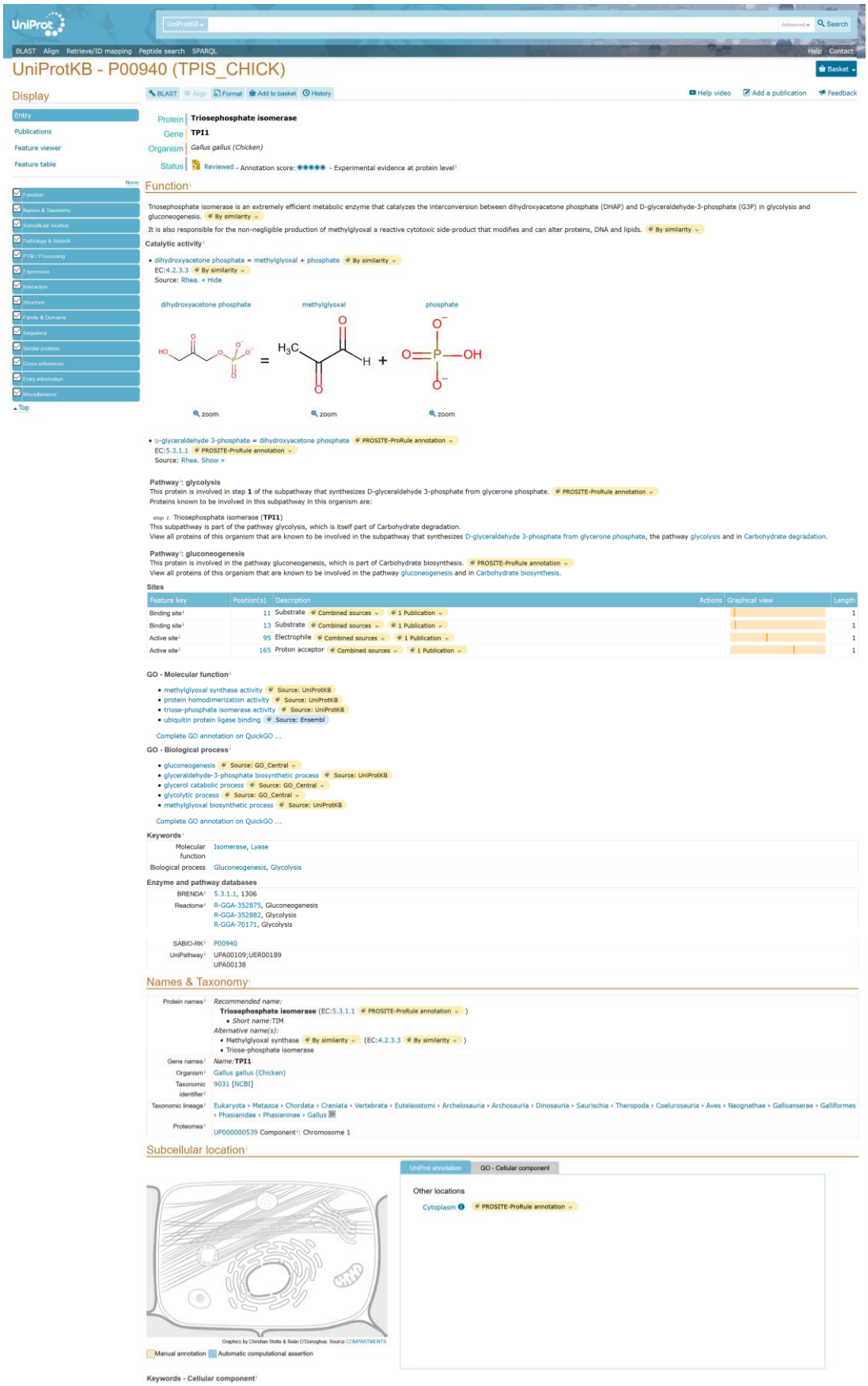


Databáza UniProt pôvodne bola založená a existovala ako databáza SwissProt a bola spravovaná v rámci Swiss Institue of Bioinformatics na tvz. Expasy serveri (http://www.expasy.ch/, resp. http://www.expasy.org/).

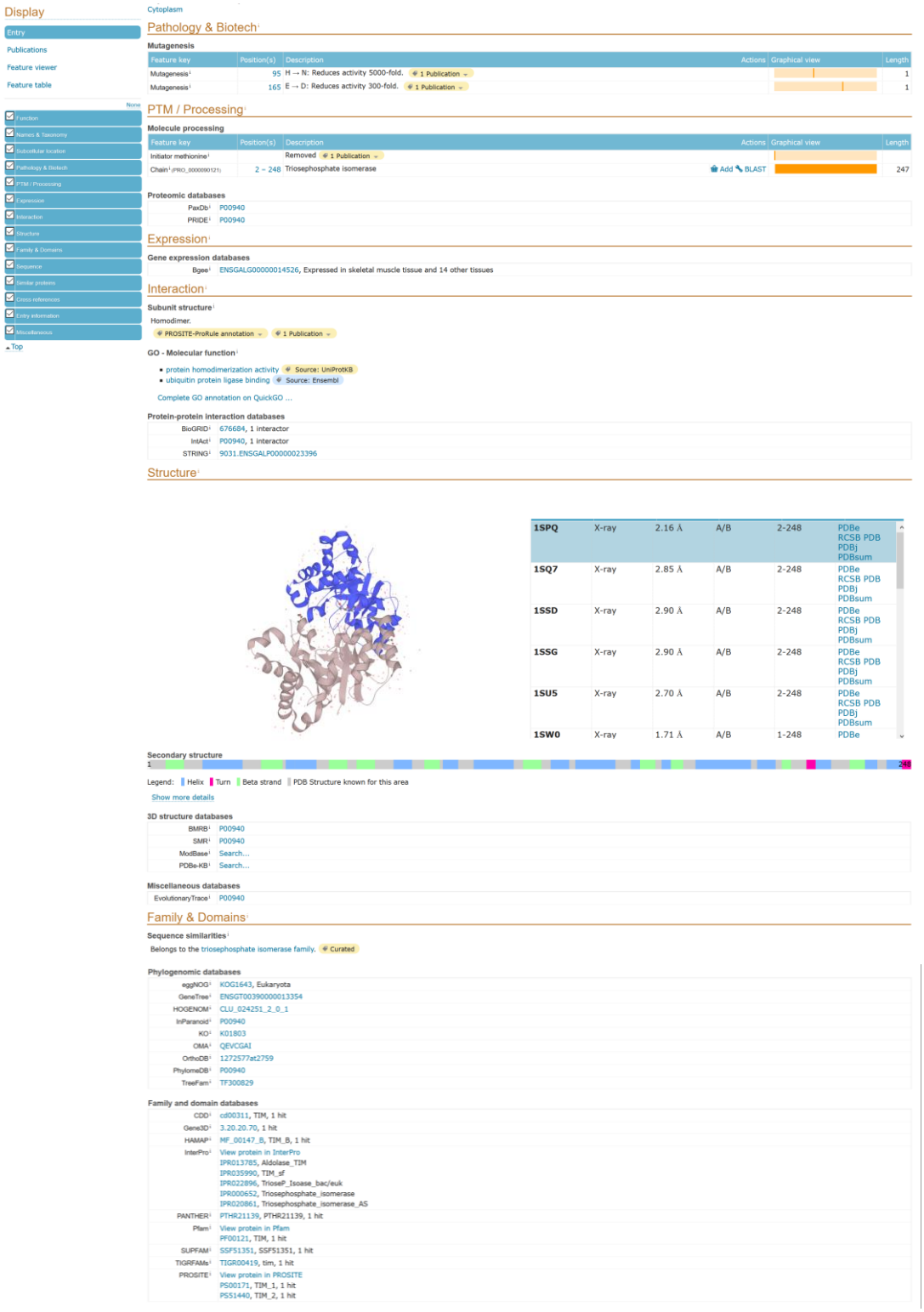
Keďže dominanciu získali databázy s nukleotidovými sekvenciami (GenBank, EMBL/ENA, DDBJ) – v dôsledku prevahy sekvenovania nukleotidov pred sekvenovaním aminokyselín – k databáze SwissProt bola vytvorená jej súčasť TrEMBL, t.j. preložená EMBL nukleotidová databáza (translated EMBL Nucleotide Database – TrEMBL). V súčasnosti pribúdajú dáta do databázy UniProt takmer výhradne cez jej súčasť TrEMBL, keďže aminokyselinové sekvencie sa nezískavajú priamo sekvenovaním proteínov, ale prekladom nukleotidových sekvencií. Treba si ale uvedomiť, že EMBL Nucleotide Database je v súčasnosti databáza ENA (European Nucleotide Archive).

Prístupové číslo sekvencie proteínu v databáze UniProt je odlišné od prístupového čísla pre tú istú sekvenciu proteínu a jeho génu z databázy GenBank (ENA a DDBJ). Prístupové čísla z nukleotidových databáz, ako aj číslo „protein_id“ z GenPept, sú však súčasťou každého záznamu v databáze UniProt v časti „Cross-references“.

UniProt záznam pre triózafosfátizomerázu (prístupové číslo P00940):



Záznam z databázy UniProt pre triózafosfátizomerázu (pokračovanie):



Display

Entry
Publications
Feature viewer
Feature table

None
Function
Names & Taxonomy
Subcellular location
Pathology & Biotech
PTM / Processing
Expression
Interaction
Structure
Family & Domains
Sequence
Similar proteins
Cross-references
Entry information
Miscellaneous

Top

Cytoplasm

Pathology & Biotech

Mutagenesis

Feature key	Position(s)	Description	Actions	Graphical view	Length
Mutagenesis	95	H → N: Reduces activity 5000-fold.	1 Publication		1
Mutagenesis	165	E → D: Reduces activity 300-fold.	1 Publication		1

PTM / Processing

Molecule processing

Feature key	Position(s)	Description	Actions	Graphical view	Length
Initiator methionine		Removed	1 Publication		
Chain (PRO_0000090121)	2 – 248	Triosephosphate isomerase	Add BLAST		247

Proteomic databases

PaxDb	P00940
PRIDE	P00940

Expression

Gene expression databases

Bgee	ENSGALG00000014526, Expressed in skeletal muscle tissue and 14 other tissues

Interaction

Subunit structure

Homodimer.

PROSITE-ProRule annotation | 1 Publication

GO - Molecular function

- protein homodimerization activity — Source: UniProtKB
- ubiquitin protein ligase binding — Source: Ensembl

Complete GO annotation on QuickGO ...

Protein-protein interaction databases

BioGRID	676684, 1 interactor
IntAct	P00940, 1 interactor
STRING	9031.ENSGALP00000023396

Structure

1SPQ	X-ray	2.16 Å	A/B	2-248	PDBe RCSB PDB PDBj PDBsum
1SQ7	X-ray	2.85 Å	A/B	2-248	PDBe RCSB PDB PDBj PDBsum
1SSD	X-ray	2.90 Å	A/B	2-248	PDBe RCSB PDB PDBj PDBsum
1SSG	X-ray	2.90 Å	A/B	2-248	PDBe RCSB PDB PDBj PDBsum
1SU5	X-ray	2.70 Å	A/B	2-248	PDBe RCSB PDB PDBj PDBsum
1SW0	X-ray	1.71 Å	A/B	1-248	PDBe

Secondary structure

1 … 248

Legend: Helix | Turn | Beta strand | PDB Structure known for this area

Show more details

3D structure databases

BMRB	P00940
SMR	P00940
ModBase	Search...
PDBe-KB	Search...

Miscellaneous databases

EvolutionaryTrace	P00940

Family & Domains

Sequence similarities

Belongs to the triosephosphate isomerase family. Curated

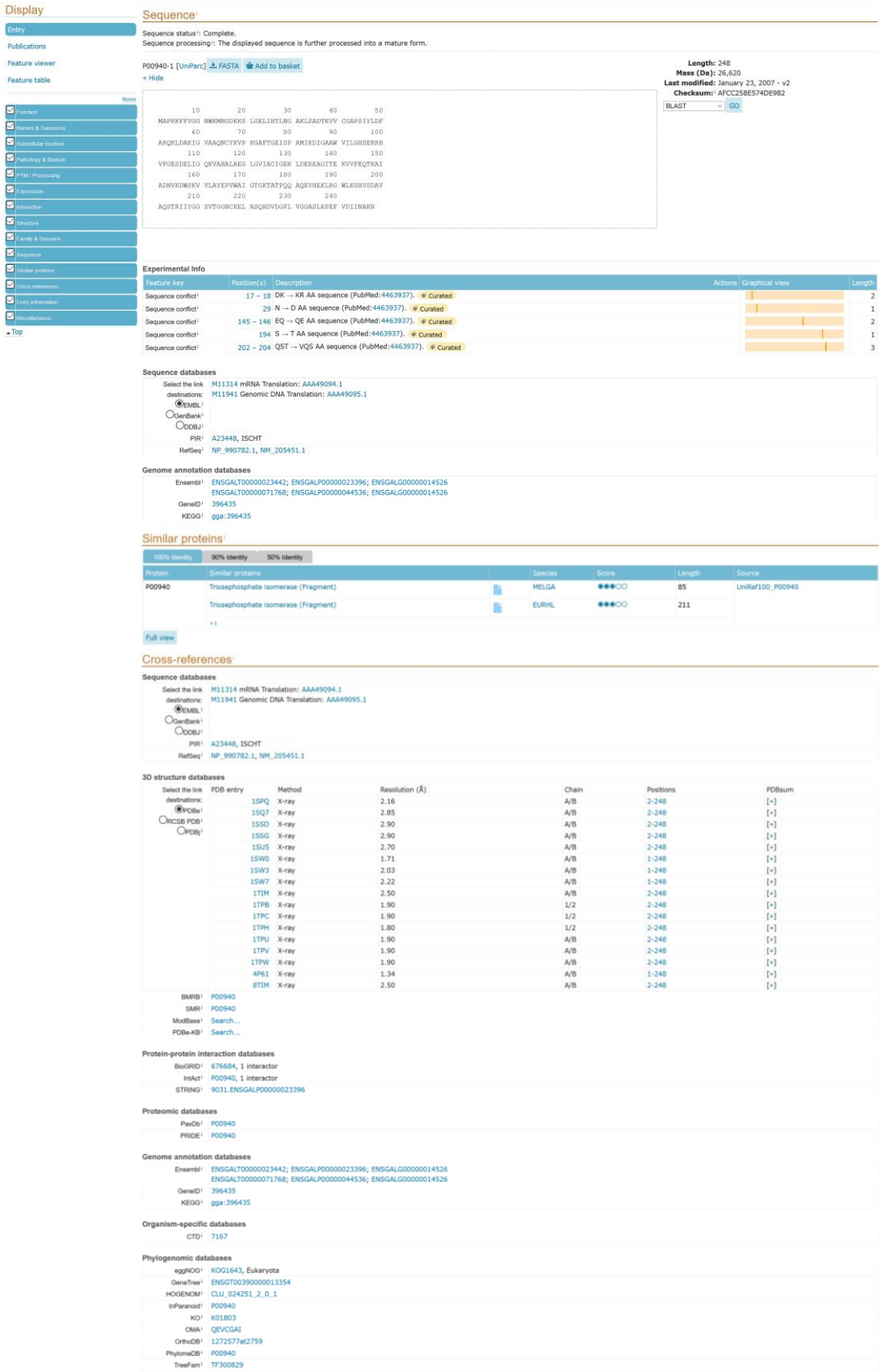
Phylogenomic databases

eggNOG	KOG1643, Eukaryota
GeneTree	ENSGT00390000013354
HOGENOM	CLU_024251_2_0_1
InParanoid	P00940
KO	K01803
OMA	QEVCGAI
OrthoDB	1272577at2759
PhylomeDB	P00940
TreeFam	TF300829

Family and domain databases

CDD	cd00311, TIM, 1 hit
Gene3D	3.20.20.70, 1 hit
HAMAP	MF_00147_B, TIM_B, 1 hit
InterPro	View protein in InterPro IPR013785, Aldolase_TIM IPR035990, TIM_sf IPR022896, TrioseP_Isoase_bac/euk IPR000652, Triosephosphate_isomerase IPR020861, Triosephosphate_isomerase_AS
PANTHER	PTHR21139, PTHR21139, 1 hit
Pfam	View protein in Pfam PF00121, TIM, 1 hit
SUPFAM	SSF51351, SSF51351, 1 hit
TIGRFAMs	TIGR00419, tim, 1 hit
PROSITE	View protein in PROSITE PS00171, TIM_1, 1 hit PS51440, TIM_2, 1 hit

Záznam z databázy UniProt pre triózafosfátizomerázu (pokračovanie):



Display

Entry
Publications
Feature viewer
Feature table

Function
Names & Taxonomy
Subcellular location
Pathology & Biotech
PTM / Processing
Expression
Interaction
Structure
Family & Domains
Sequence
Similar proteins
Cross-references
Entry information
Miscellaneous

Top

Sequence

Sequence status: Complete.
Sequence processing: The displayed sequence is further processed into a mature form.

P00940-1 [UniParc] FASTA Add to basket
Hide

Length: 248
Mass (Da): 26,620
Last modified: January 23, 2007 - v2
Checksum: AFCC258E574DE982

BLAST GO

```
        10         20         30         40         50
MAPRKFFVGG NWKMNGDKKS LGKLIHTLNG AKLSADTEVV CGAPSIYLDF
        60         70         80         90        100
ARQKLDAKIG VAAQNCYKVP KGAFTGEISP AMIKDIGAAW VILGHSERRH
       110        120        130        140        150
VFGESDELIG QKVAHALAEG LGVIACIGEK LDEREAGITE KVVFEQTKAI
       160        170        180        190        200
ADNVKDWSKV VLAYEPVWAI GTGKTATPQQ AQEVHEKLRG WLKSHVSDAV
       210        220        230        240
AQSTRIIYGG SVTGGNCKEL ASQHDVDGFL VGGASLKPEF VDIINAKH
```

Experimental info

Feature key	Position(s)	Description	Actions	Graphical view	Length
Sequence conflict	17 – 18	DK → KR AA sequence (PubMed:4463937). Curated			2
Sequence conflict	29	N → D AA sequence (PubMed:4463937). Curated			1
Sequence conflict	145 – 146	EQ → QE AA sequence (PubMed:4463937). Curated			2
Sequence conflict	194	S → T AA sequence (PubMed:4463937). Curated			1
Sequence conflict	202 – 204	QST → VQS AA sequence (PubMed:4463937). Curated			3

Sequence databases

Select the link destinations: EMBL GenBank DDBJ
M11314 mRNA Translation: AAA49094.1
M11941 Genomic DNA Translation: AAA49095.1
PIR: A23448, ISCHT
RefSeq: NP_990782.1, NM_205451.1

Genome annotation databases

Ensembl: ENSGALT00000023442; ENSGALP00000023396; ENSGALG00000014526
ENSGALT00000071768; ENSGALP00000044536; ENSGALG00000014526
GeneID: 396435
KEGG: gga:396435

Similar proteins

100% Identity | 90% Identity | 50% Identity

Protein	Similar proteins		Species	Score	Length	Source
P00940	Triosephosphate isomerase (Fragment)		MELGA	●●●○○	85	UniRef100_P00940
	Triosephosphate isomerase (Fragment)		EURHL	●●●○○	211	

+1

Full view

Cross-references

Sequence databases

Select the link destinations: EMBL GenBank DDBJ
M11314 mRNA Translation: AAA49094.1
M11941 Genomic DNA Translation: AAA49095.1
PIR: A23448, ISCHT
RefSeq: NP_990782.1, NM_205451.1

3D structure databases

Select the link destinations: PDBe RCSB PDB PDBj

PDB entry	Method	Resolution (Å)	Chain	Positions	PDBsum
1SPQ	X-ray	2.16	A/B	2-248	[+]
1SQ7	X-ray	2.85	A/B	2-248	[+]
1SSD	X-ray	2.90	A/B	2-248	[+]
1SSG	X-ray	2.90	A/B	2-248	[+]
1SU5	X-ray	2.70	A/B	2-248	[+]
1SW0	X-ray	1.71	A/B	1-248	[+]
1SW3	X-ray	2.03	A/B	1-248	[+]
1SW7	X-ray	2.22	A/B	1-248	[+]
1TIM	X-ray	2.50	A/B	2-248	[+]
1TPB	X-ray	1.90	1/2	2-248	[+]
1TPC	X-ray	1.90	1/2	2-248	[+]
1TPH	X-ray	1.80	1/2	2-248	[+]
1TPU	X-ray	1.90	A/B	2-248	[+]
1TPV	X-ray	1.90	A/B	2-248	[+]
1TPW	X-ray	1.90	A/B	2-248	[+]
4P61	X-ray	1.34	A/B	1-248	[+]
8TIM	X-ray	2.50	A/B	2-248	[+]

BMRB: P00940
SMR: P00940
ModBase: Search...
PDBe-KB: Search...

Protein-protein interaction databases

BioGRID: 676684, 1 interactor
IntAct: P00940, 1 interactor
STRING: 9031.ENSGALP00000023396

Proteomic databases

PaxDb: P00940
PRIDE: P00940

Genome annotation databases

Ensembl: ENSGALT00000023442; ENSGALP00000023396; ENSGALG00000014526
ENSGALT00000071768; ENSGALP00000044536; ENSGALG00000014526
GeneID: 396435
KEGG: gga:396435

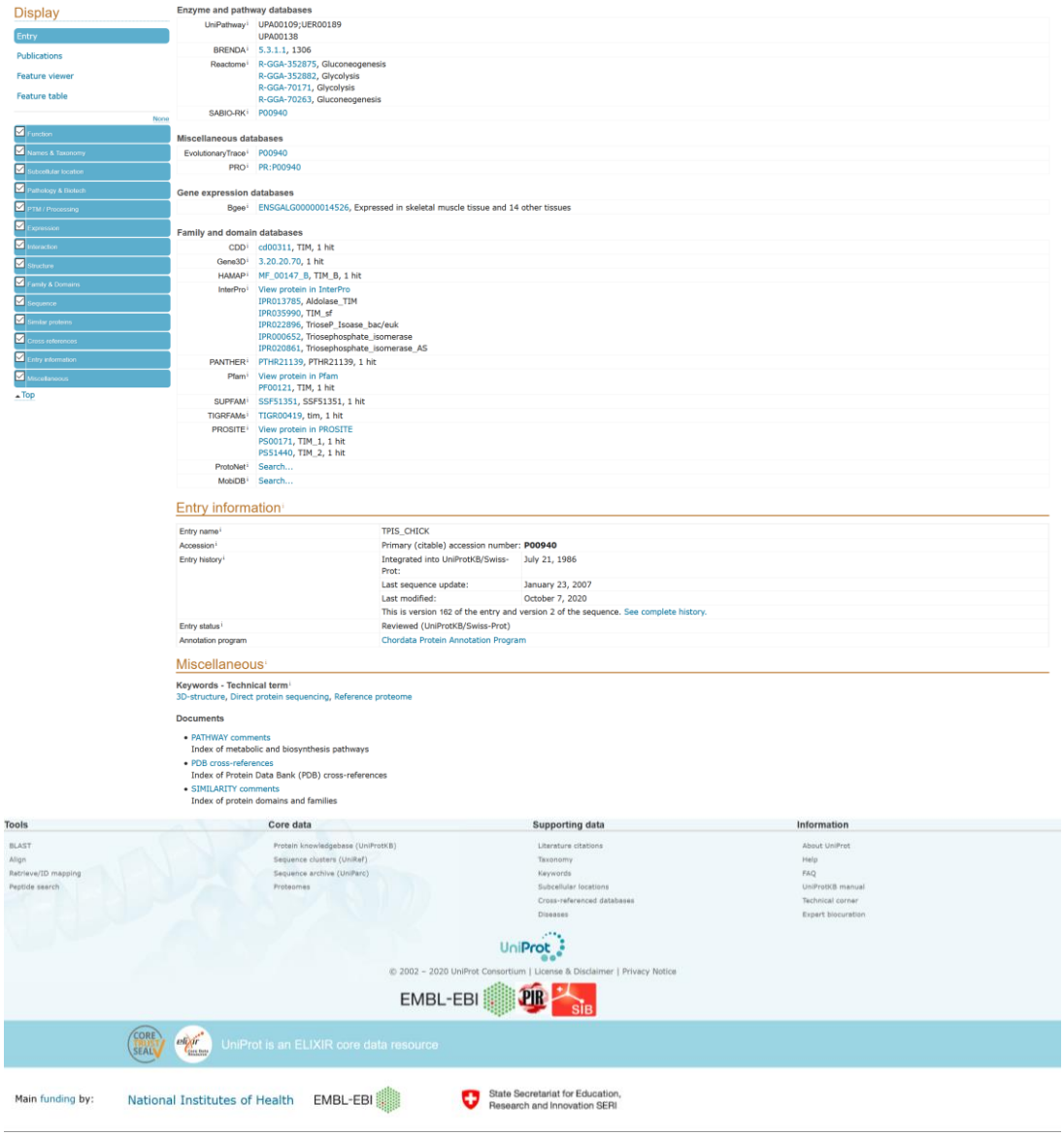
Organism-specific databases

CTD: 7167

Phylogenomic databases

eggNOG: KOG1643, Eukaryota
GeneTree: ENSGT00390000013354
HOGENOM: CLU_024251_2_0_1
InParanoid: P00940
KO: K01803
OMA: QEVCGAI
OrthoDB: 1272577at2759
PhylomeDB: P00940
TreeFam: TF300829

Záznam z databázy UniProt pre triózafosfátizomerázu (dokončenie):



Pre prácu so sekvenciami – aj nukleotidovými, aj aminokyselinovými – sa používa ich zápis v tzv. FASTA formáte v textovom súbore. V ňom je každá sekvencia uvedená v prvom riadku znakom „>“, za ktorým je jej názov (napr. vhodná skratka), pričom samotná sekvencia začína na nasledujúcom riadku:

```
>sp|P00940|TPIS_CHICK Triosephosphate isomerase OS=Gallus gallus OX=9031 GN=TPI1 PE=1 SV=2
MAPRKFFVGGNWKMNGDKKSLGELIHTLNGAKLSADTEVVCGAPSIYLDFARQKLDAKIG
VAAQNCYKVPKGAFTGEISPAMIKDIGAAWVILGHSERRHVFGESDELIGQKVAHALAEG
LGVIACIGEKLDEREAGITEKVVFEQTKAIADNVKDWSKVVLAYEPVWAIGTGKTATPQQ
AQEVHEKLRGWLKSHVSDAVAQSTRIIYGGSVTGGNCKELASQHDVDGFLVGGASLKPEF
VDIINAKH
```

3. Základy bioinformatickej – *in silico* – analýzy proteínov

Základy práce pri bioinformatickej analýze sekvencií proteínov, tzv. *in silico* analýze, možno ilustrovať na nasledujúcom vzorovom príklade. Ide o porovnanie 7 sekvencií enzýmu triózafosfátizomeráza, pochádzajúcich z rôznych zástupcov baktérií, archeónov a eukaryotov (tab. 3.1).

Tabuľka 3.1. Zoznam študovaných triózafosfátizomeráz.

Č.	Zdroj	Skratka	UniProt	Dĺžka
	Bacteria			
1	*Escherichia coli*	Escco	P0A858	255
2	*Nostoc* sp. PCC 7120	Nossp	Q8YP17	241
	Archaea			
3	*Pyrococcus furiosus*	Pyrfu	P62002	228
4	*Sulfolobus solfataricus*	Sulso	Q97VM8	227
	Eucarya			
5	*Saccharomyces cerevisiae*	Sacce	P00942	248
6	*Arabidopsis thaliana*	Arath	Q9SKP6	255 (61-315)
7	*Homo sapiens* (izoforma 1)	Homsa	P60174	249

Celý postup práce možno zosumarizovať ako zadanie v nasledujúcich krokoch:

(1) Vytvorte si priečinok „TIM“.

(2) Všetkých 7 sekvencií TIM zhromaždite z databázy UniProt do vstupného súboru („TIM.txt“) vhodného pre program Clustal-Omega.

(3) Na EBI serveri zrovnajte sekvencie TIM v programe Clustal-Omega:

http://www.ebi.ac.uk/Tools/msa/clustalo/

a získajte dva súbory: (i) formát ALIGNMENT – „Clustal with character counts“ (súbor: „TIM_aln.txt“); a (ii) formát Pearson/FASTA (súbor: „TIM_fas.txt“);); dodržte vstupné poradie sekvencií (input order).

(4) Súbor „TIM_aln.txt“ otvorte v editovacom programe (napr. MS-Word) a v zrovnaných sekvenciách zvýraznite žltým podfarbením aminokyselinové zvyšky aktívneho miesta (Asn10, Lys12, His95 a Glu165 v sekvencii triózafosfátizomerázy zo *Saccharomyces cerevisaiae*); súbor uložte ako dokument TIM_aln.doc.

(5) Na EBI serveri v rámci nástroja Simple Phylogeny:

http://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/

vypočítajte súbory pre dva evolučné stromy pre TIM: (i) s uvažovaním medzier v sekvenciách – podmienka „Exclude gaps“ je vypnutá: „off“ (súbor „TIM_off.txt“); (ii) a s ich ignorovaním – podmienka „Exclude gaps“ je zapnutá: „on“ (súbor „TIM_on.txt“). Oba súbory počítajte na základe finálneho súboru zrovnania sekvencií vo formáte Pearson/FASTA – „TIM_fas.txt“ (ktorý je obsahovo zhodný so súborom zrovnania vo formáte „Clustal with character counts“ – „TIM_aln.txt“).

(6) Vypočítajte hodnoty konsenzuálnej dĺžky (CL), sekvenčnej identity (SI) a sekvenčnej podobnosti (SS).

(7) Vypočítané evolučné stromy zobrazte v programe iTOL (Interactive Tree of Life; https://itol.embl.de/) – napr. v režime „display mode“ ako „circular“ (t.j. kruhový typ stromu), vložte ako exportované obrázky (napr. PNG) do súboru „TIM_aln.doc“, porovnajte a zapíšte diskusiu celej práce.

Príprava vstupného súboru

Po vytvorení priečinku („TIM“), do ktorého sa budú – kvôli prehľadnosti – ukladať všetky údaje, je potrebné vytvoriť si tzv. vstupný súbor so sekvenciami. Tento súbor („TIM.txt“) je súbor textového typu, pričom všetky sekvencie v ňom uložené musia byť v tzv. FASTA formáte (obr. 3.1). Pre jednoduché získanie sekvencií sú v tomto konkrétnom prípade udané prístupové čísla (Accession Nos.) z databázy UniProt pre všetkých 7 študovaných triózafosfátizomeráz (tab. 3.1). Sekvencie by samozrejme mohli byť získané aj z nukleotidovej databázy (GenBank/ENA/DDBJ), ak by boli zadané prístupové čísla z týchto nukleotidových databáz, prípadne tzv. GenPept „protein_id“ čísla (pre súbory obsahujúce iba aminokyselinové sekvencie). V každom prípade – akýmkoľvek spôsobom sa sekvencie získajú – vo vstupnom súbore sa musia uložiť vo FASTA formáte a súbor musí byť textový súbor, t.j. obsahuje iba text (nemôže to byť napr. súbor programu MS-Word uložený ako „dokument“ – súbor typu „DOC“).

Zrovnávanie sekvencií

V ďalšom kroku je potrebné vykonať samotné zrovnanie sekvencií. Na to je možné využiť rôzne zrovnávacie programy, väčšinou voľne dostupné na internete. Jedným z nich je aj program Clustal-Omega.

```
>Escco
MRHPLVMGNWKLNGSRHMVHELVSNLRKELAGVAGCAVAIAPPEMYIDMAKREAEGSHIM
LGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSERRTYHKESDELIAKKFAVLKEQG
LTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWAIGTGKSATP
AQAQAVHKFIRDHIAKVDANIAEQVIIQYGGSVNASNAAELFAQPDIDGALVGGASLKAD
AFAVIVKAAEAAKQA

>Nossp
MRKIVIAGNWKMFKTQAESQEFLKEFLPALEETPQEREVLLCVPFTDLAILSQSLHGSLV
QLGAQNVHWAENGAYTGEISGPMLTEIGVRYVIVGHSERRQFFGETDETVNLRLQAAQKY
GLTPILCVGETKQQRDSGETESLIVSQLDKDLINVDQTNLVIAYEPIWAIGTGDTCETTE
ANRVIGLIRSQLKNSDVPIQYGGSVKPNNIDEIMAQPEIDGVLVGGASLEAASFARIVNY
L

>Pyrfu
MAKLKEPIIAINFKTYIEATGKRALEIAKAAEKVYKETGVTIVVAPQLVDLRMIAESVEI
PVFAQHIDPIKPGSHTGHVLPEAVKEAGAVGTLLNHSENRMILADLEAAIRRAEEVGLMT
MVCSNNPAVSAAVAALNPDYVAVEPPELIGTGIPVSKAKPEVITNTVELVKKVNPEVKVL
CGAGISTGEDVKKAIELGTVGVLLASGVTKAKDPEKAIWDLVSGIIKE

>Sulso
MKPPIIIINFKAYENSFGDKAVNLGKKIEKISKEYSVEIILSTPATMIYRMSQEVDLPIY
AEHVDPVPLGAFTGAILPEMVKDAGAKGTLINHSERRLRADEIDDVLKRTKKLGLKSILC
VDRYELVYPFSLLKPDAILIEPPELIGTGVSVSKAKPEVITRAVDEIRKSEGIYLIAGAG
ITTGEDVYKALKLGAHGIGVASAVMKAKEPEKVVEDFITSALRAISS

>Sacce
MARTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVKKPQVT
VGAQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSERRSYFHEDDKFIADKTKFALGQG
VGVILCIGETLEEKKAGKTLDVVERQLNAVLEEVKDWTNVVVAYEPVWAIGTGLAATPED
AQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKPEF
VDIINSRN

>Arath
AGSGKFFVGGNWKCNGTKDSIAKLISDLNSATLEADVDVVVSPPFVYIDQVKSSLTDRID
ISGQNSWVGKGGAFTGEISVEQLKDLGCKWVILGHSERRHVIGEKDEFIGKKAAYALSEG
LGVIACIGEKLEEREAGKTFDVCFAQLKAFADAVPSWDNIVVAYEPVWAIGTGKVASPQQ
AQEVHVAVRGWLKKNVSEEVASKTRIIYGGSVNGGNSAELAKEEDIDGFLVGGASLKGPE
FATIVNSVTSKKVAA

>Homsa
MAPSRKFFVGGNWKMNGRKQSLGELIGTLNAAKVPADTEVVCAPPTAYIDFARQKLDPKI
AVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGESDELIGQKVAHALAE
GLGVIACIGEKLDEREAGITEKVVFEQTKVIADNVKDWSKVVLAYEPVWAIGTGKTATPQ
QAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGFLVGGASLKPE
FVDIINAKQ
```

Obr. 3.1. Vstupný súbor so sekvenciami pripravenými na zrovnávanie. V tomto konkrétnom prípade – podľa zadania – napr. „TIM.txt“. Je to zápis v tzv. FASTA formáte; typ súboru je textový (t.j. iba text). Každá sekvencia je uvedená v prvom riadku znakom „>“, za ktorým sa nachádza jej názov (napr. vhodná skratka, ale môže to byť aj napr. prístupové číslo z databázy, apod.), pričom samotná sekvencia začína na nasledujúcom riadku. To sa opakuje pre všetky sekvencie.

Program Clustal-Omega je využiteľný online ako „nástroj“ cez web-stránky Európskeho ústavu pre bioinformatiku (EBI). Prostredie programu Clustal-Omega je užívateľsky prívetivé a na obsluhu jednoduché.

Z jedného a toho istého vstupného súboru (v zmysle zadania; obr. 3.1) sa pripravia dva súbory so zrovnanými sekvenciami:

(i) súbor typu ALIGNMENT, v ktorom sú sekvencie zrovnané v blokoch po 60 pozíciií, pričom na konci každého bloku sú udané číselné pozície príslušných aminokyselinových zvyškov z jednotlivých sekvencií (súbor „TIM_aln.txt)“ – obr. 3.2;

(ii) súbor typu Pearson/FASTA, v ktorom sú sekvencie zrovnané vždy každá jedna samostatne a celá v jednom kuse; tento súbor je podobný vstupnému súboru (je to FASTA formát sekvencií), ale s tým rozdielom, že sekvencie sú už zrovnané, t.j. zarovnané na rovnakú (tzv. konsenzuálnu) dĺžku a sú v nich povkladané medzery (súbor: „TIM_fas.txt“) – obr. 3.3.

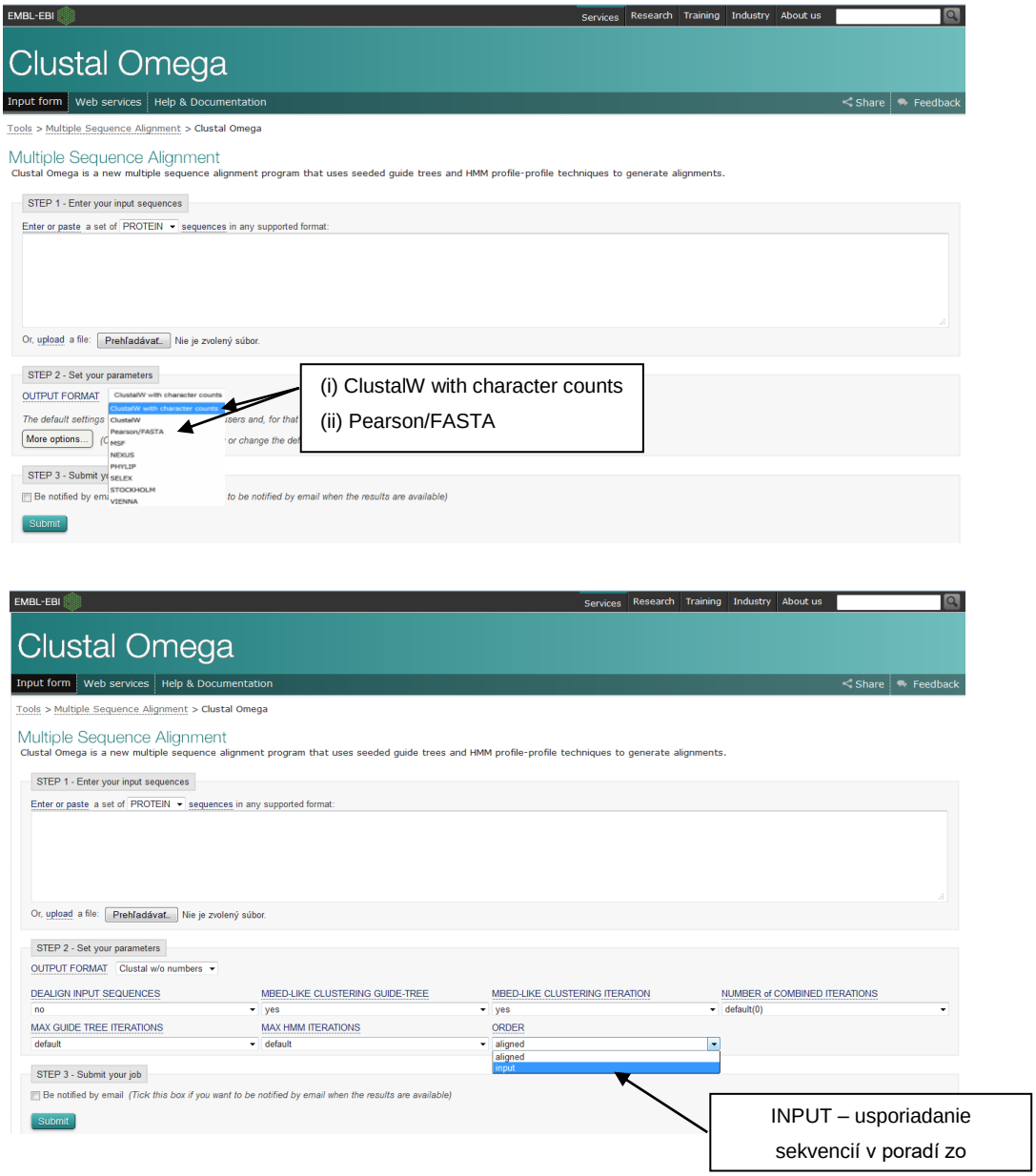
V prvom prípade treba v programe zvoliť podmienku cez „Output format“ ako „Clustal with character counts“, kým v druhom prípade je to podmienka „Pearson/FASTA“. Oba súbory sú obsahovo identické, t.j. rozdiel medzi nimi je len formálny – predstavujú rozdielne formy zápisu. Cez podmienku „More options“ a „Order“ je vhodné zvoliť „input order” (t.j. nie “aligned”), aby sa v zrovnaní dodržalo vstupné poradie sekvencií.

Súbor zrovnania typu „ALIGNMENT“ je vhodnejší pre používateľa, napr. pre ďalšiu analýzu a prípadnú editáciu, ako aj pre prípravu obrázkov. Súbor zrovnania vo formáte „Pearson/FASTA“ potom slúži v ďalšej práci ako vstupný súbor pre výpočet evolučného stromu, resp. evolučných stromov.

Pri práci – editácii – zrovnania vo formáte „ALIGNMENT“, napr. pri identifikácii a zvýrazňovaní určitých špecifických čŕt jednotlivých sekvencií – ako môžu byť funkčne dôležité aminokyselinové zvyšky alebo nejaké konzervované sekvenčné regióny, apod. – je potrebné sa správne v zrovnaní orientovať. Treba napr. rozlišovať medzi pozíciami v zrovnaní a pozíciami jednotlivých zvyškov v príslušných sekvenciách. Taktiež treba – napr. ak sú zadané nejaké konkrétne aminokyselinové zvyšky – vždy tieto pozície správne identifikovať v sekvencii z daného zdroja (organizmu), aby nedošlo k chybám

a zavádzajúcim interpretáciám. Toto sa týka zvyškov Asn10, Lys12, His95 a Glu165 v sekvencii triózafosfátizomerázy zo *Saccharomyces cerevisaiae* – pre konkrétny prípad 7 študovaných triózafosfátizomeráz (tab. 3.1) – zvyšky treba identifikovať s ohľadom na dané aminokyseliny (t.j. N, K, H a E), ich pozície v sekvencii (t.j. nie pozície v zrovnaní) a s ohľadom na ich zdroj (*Saccharomyces cerevisiae*). Až potom je možné pristúpiť k prípadnému zvýrazneniu korešpondujúcich zvyškov v danej pozícii pre všetky ostatné zdroje v rámci celého študovaného súboru (obr. 3.4).

Internetové prostredie programu Clustal-Omega na EBI serveri:
http://www.ebi.ac.uk/Tools/msa/clustalo/



```
Escco   ---MRHPLVMGNWKLNGSRH-----MVHELVSNLRK-ELAGVAGCAVAIAPPEMYIDMAK    51
Nossp   ---MRKIVIAGNWKMFKTQA-----ESQEFLKEFLPALEETPQEREVLLCVPFTDLAILS    52
Pyrfu   MAKLKEPIIAINFKTYIEATGKRALEIAKAAEKVYK-----ETGVTIVVAPQLVDLRMIA    55
Sulso   ---MKPPIIIINFKAYENSFGDKAVNLGKKIEKISK-----EYSVEIILSTPATMIYRMS    52
Sacce   --MARTFFVGGNFKLNGSKQ-----SIKEIVERLNT-ASI-PENVEVVICPPATYLDYSV    51
Arath   -AGSGKFFVGGNWKCNGTKD-----SIAKLISDLNS-ATL-EADVDVVVSPPFVYIDQVK    52
Homsa   MAPSRKFFVGGNWKMNGRKQ-----SLGELIGTLNA-AKV-PADTEVVCAPPTAYIDFAR    53
              .:  *:*             :    .           :  .     :

Escco   REAEGSHIMLGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSERRTYHKESDELIAK   111
Nossp   QSLHGSLVQLGAQNVHWAENGAYTGEISGPMLTEIGVRYVIVGHSERRQFFGETDETVNL   112
Pyrfu   ESVE---IPVFAQHIDPIKPGSHTGHVLPEAVKEAGAVGTLLNHSENRMILADLEAAIR-   111
Sulso   QEVD---LPIYAEHVDPVPLGAFTGAILPEMVKDAGAKGTLINHSERRLRADEIDDVLK-   108
Sacce   SLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSERRSYFHEDDKFIAD   111
Arath   SSL-TDRIDISGQNSWVGKGGAFTGEISVEQLKDLGCKWVILGHSERRHVIGEKDEFIGK   111
Homsa   QKL-DPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGESDELIGQ   112
              : : .:.      *:.**      :.: *    :: ***.*    : :  :

Escco   KFAVLKEQGLTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWA   171
Nossp   RLQAAQKYGLTPILCVGETKQQRDSGETESLIVSQLDKDLIN---VDQTNLVIAYEPIWA   169
Pyrfu   ---RAEEVGLMTMVCSNNPAVS--------AA------------VAALNPDYVAVEPPEL   148
Sulso   ---RTKKLGLKSILCVDRYELV--------YP------------FSLLKPDAILIEPPEL   145
Sacce   KTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEE--VKDWTNVVVAYEPVWA   169
Arath   KAAYALSEGLGVIACIGEKLEEREAGKTFDVCFAQLKAFADA--VPSWDNIVVAYEPVWA   169
Homsa   KVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKVIADN--VKDWSKVVLAYEPVWA   170
              *:  : *  .                                 :  **

Escco   IGTGKSATPAQAQAVHKFIRDHIAK-VDANIAEQVIIQYGGSVNASNAAELFAQPDIDGA   230
Nossp   IGTGDTCETTEANRVIGLIRSQLKN-------SDVPIQYGGSVKPNNIDEIMAQPEIDGV   222
Pyrfu   IGTGIPVSKAKPEVITNTVELVKKVN------PEVKVLCGAGISTGEDVKKAIELGTVGV   202
Sulso   IGTGVSVSKAKPEVITRAVDEIRKS-------EGIYLIAGAGITTGEDVYKALKLGAHGI   198
Sacce   IGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGF   229
Arath   IGTGKVASPQQAQEVHVAVRGWLKKNVSEEVASKTRIIYGGSVNGGNSAELAKEEDIDGF   229
Homsa   IGTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGF   230
        ****      . : :   :                 :  *.. .         :    *

Escco   LVGGASLKADAFAVIVKAAEAAKQA----   255
Nossp   LVGGASLEAASFARIVNYL----------   241
Pyrfu   LLASGVTKAKDPEKAIWDLVSGIIKE---   228
Sulso   GVASAVMKAKEPEKVVEDFITSALRAISS   227
Sacce   LVGGASLK-PEFVDIINSRN---------   248
Arath   LVGGASLKGPEFATIVNSVTSKKVAA---   255
Homsa   LVGGASLK-PEFVDIINAKQ---------   249
         :...  :       :
```

Obr. 3.2. Zrovnanie vo formáte ALIGNMENT – „Clustal with character counts“, t.j. zrovnanie v blokoch s počítaním zvyškov. V tomto konkrétnom prípade – podľa zadania – napr. „TIM_aln.txt“. Sekvencie sú zrovnané v blokoch po 60 pozícií, pričom sú v nich povkladané medzery, aby bola maximalizovaná ich podobnosť. Identické a podobné korešpondujúce zvyšky sú zvýraznené symbolmi „*“, resp. „:“ a „.“. Čísla vpravo udávajú pozície koncových aminokyselín v príslušnom bloku a v príslušnej sekvencii.

```
>Escco
---MRHPLVMGNWKLNGSRH-----MVHELVSNLRK-ELAGVAGCAVAIAPPEMYIDMAK
REAEGSHIMLGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSERRTYHKESDELIAK
KFAVLKEQGLTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWA
IGTGKSATPAQAQAVHKFIRDHIAK-VDANIAEQVIIQYGGSVNASNAAELFAQPDIDGA
LVGGASLKADAFAVIVKAAEAAKQA----

>Nossp
---MRKIVIAGNWKMFKTQA-----ESQEFLKEFLPALEETPQEREVLLCVPFTDLAILS
QSLHGSLVQLGAQNVHWAENGAYTGEISGPMLTEIGVRYVIVGHSERRQFFGETDETVNL
RLQAAQKYGLTPILCVGETKQQRDSGETESLIVSQLDKDLIN---VDQTNLVIAYEPIWA
IGTGDTCETTEANRVIGLIRSQLKN-------SDVPIQYGGSVKPNNIDEIMAQPEIDGV
LVGGASLEAASFARIVNYL----------

>Pyrfu
MAKLKEPIIAINFKTYIEATGKRALEIAKAAEKVYK-----ETGVTIVVAPQLVDLRMIA
ESVE---IPVFAQHIDPIKPGSHTGHVLPEAVKEAGAVGTLLNHSENRMILADLEAAIR-
---RAEEVGLMTMVCSNNPAVS--------AA------------VAALNPDYVAVEPPEL
IGTGIPVSKAKPEVITNTVELVKKVN------PEVKVLCGAGISTGEDVKKAIELGTVGV
LLASGVTKAKDPEKAIWDLVSGIIKE---

>Sulso
---MKPPIIIINFKAYENSFGDKAVNLGKKIEKISK-----EYSVEIILSTPATMIYRMS
QEVD---LPIYAEHVDPVPLGAFTGAILPEMVKDAGAKGTLINHSERRLRADEIDDVLK-
---RTKKLGLKSILCVDRYELV--------YP------------FSLLKPDAILIEPPEL
IGTGVSVSKAKPEVITRAVDEIRKS-------EGIYLIAGAGITTGEDVYKALKLGAHGI
GVASAVMKAKEPEKVVEDFITSALRAISS

>Sacce
--MARTFFVGGNFKLNGSKQ-----SIKEIVERLNT-ASI-PENVEVVICPPATYLDYSV
SLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSERRSYFHEDDKFIAD
KTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEE--VKDWTNVVVAYEPVWA
IGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGF
LVGGASLK-PEFVDIINSRN---------

>Arath
-AGSGKFFVGGNWKCNGTKD-----SIAKLISDLNS-ATL-EADVDVVVSPPFVYIDQVK
SSL-TDRIDISGQNSWVGKGGAFTGEISVEQLKDLGCKWVILGHSERRHVIGEKDEFIGK
KAAYALSEGLGVIACIGEKLEEREAGKTFDVCFAQLKAFADA--VPSWDNIVVAYEPVWA
IGTGKVASPQQAQEVHVAVRGWLKKNVSEEVASKTRIIYGGSVNGGNSAELAKEEDIDGF
LVGGASLKGPEFATIVNSVTSKKVAA---

>Homsa
MAPSRKFFVGGNWKMNGRKQ-----SLGELIGTLNA-AKV-PADTEVVCAPPTAYIDFAR
QKL-DPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGESDELIGQ
KVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKVIADN--VKDWSKVVLAYEPVWA
IGTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGF
LVGGASLK-PEFVDIINAKQ---------
```

Obr. 3.3. Zrovnanie sekvencií vo formáte Pearson/FASTA. V tomto konkrétnom prípade – podľa zadania – napr. „TIM_fas.txt“. Je to zápis v tzv. FASTA formáte, t.j. rovnakom, ako sa používa pre prípravu vstupného súboru (obr. 3.1). Rozdiel je v tom, že tu už sú sekvencie zrovnané, t.j. sú v nich povkladané medzery v snahe maximalizovať podobnosť medzi sekvenciami. Je to formálne odlišný, ale obsahovo identický súbor so zrovnaním vo formáte ALIGNMENT (obr. 3.2).

```
Escco   ---MRHPLVMGNWKLNGSRH-----MVHELVSNLRK-ELAGVAGCAVAIAPPEMYIDMAK    51
Nossp   ---MRKIVIAGNWKMFKTQA-----ESQEFLKEFLPALEETPQEREVLLCVPFTDLAILS    52
Pyrfu   MAKLKEPIIAINFKTYIEATGKRALEIAKAAEKVYK-----ETGVTIVVAPQLVDLRMIA    55
Sulso   ---MKPPIIIINFKAYENSFGDKAVNLGKKIEKISK-----EYSVEIILSTPATMIYRMS    52
Sacce   --MARTFFVGGNFKLNGSKQ-----SIKEIVERLNT-ASI-PENVEVVICPPATYLDYSV    51
Arath   -AGSGKFFVGGNWKCNGTKD-----SIAKLISDLNS-ATL-EADVDVVVSPPFVYIDQVK    52
Homsa   MAPSRKFFVGGNWKMNGRKQ-----SLGELIGTLNA-AKV-PADTEVVCAPPTAYIDFAR    53
                .:  *:*             :    .           :  .     :

Escco   REAEGSHIMLGAQNVDLNLSGAFTGETSAAMLKDIGAQYIIIGHSERRTYHKESDELIAK   111
Nossp   QSLHGSLVQLGAQNVHWAENGAYTGEISGPMLTEIGVRYVIVGHSERRQFFGETDETVNL   112
Pyrfu   ESVE---IPVFAQHIDPIKPGSHTGHVLPEAVKEAGAVGTLLNHSENRMILADLEAAIR-   111
Sulso   QEVD---LPIYAEHVDPVPLGAFTGAILPEMVKDAGAKGTLINHSERRLRADEIDDVLK-   108
Sacce   SLVKKPQVTVGAQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSERRSYFHEDDKFIAD   111
Arath   SSL-TDRIDISGQNSWVGKGGAFTGEISVEQLKDLGCKWVILGHSERRHVIGEKDEFIGK   111
Homsa   QKL-DPKIAVAAQNCYKVTNGAFTGEISPGMIKDCGATWVVLGHSERRHVFGESDELIGQ   112
               : : .:.      *:.**      :.: *    :: ***.*    : :  :

Escco   KFAVLKEQGLTPVLCIGETEAENEAGKTEEVCARQIDAVLKTQGAAAFEGAVIAYEPVWA   171
Nossp   RLQAAQKYGLTPILCVGETKQQRDSGETESLIVSQLDKDLIN---VDQTNLVIAYEPIWA   169
Pyrfu   ---RAEEVGLMTMVCSNNPAVS--------AA------------VAALNPDYVAVEPPEL   148
Sulso   ---RTKKLGLKSILCVDRYELV--------YP------------FSLLKPDAILIEPPEL   145
Sacce   KTKFALGQGVGVILCIGETLEEKKAGKTLDVVERQLNAVLEE--VKDWTNVVVAYEPVWA   169
Arath   KAAYALSEGLGVIACIGEKLEEREAGKTFDVCFAQLKAFADA--VPSWDNIVVAYEPVWA   169
Homsa   KVAHALAEGLGVIACIGEKLDEREAGITEKVVFEQTKVIADN--VKDWSKVVLAYEPVWA   170
                *:  : *  .                                    :  **

Escco   IGTGKSATPAQAQAVHKFIRDHIAK-VDANIAEQVIIQYGGSVNASNAAELFAQPDIDGA   230
Nossp   IGTGDTCETTEANRVIGLIRSQLKN-------SDVPIQYGGSVKPNNIDEIMAQPEIDGV   222
Pyrfu   IGTGIPVSKAKPEVITNTVELVKKVN------PEVKVLCGAGISTGEDVKKAIELGTVGV   202
Sulso   IGTGVSVSKAKPEVITRAVDEIRKS-------EGIYLIAGAGITTGEDVYKALKLGAHGI   198
Sacce   IGTGLAATPEDAQDIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGF   229
Arath   IGTGKVASPQQAQEVHVAVRGWLKKNVSEEVASKTRIIYGGSVNGGNSAELAKEEDIDGF   229
Homsa   IGTGKTATPQQAQEVHEKLRGWLKSNVSDAVAQSTRIIYGGSVTGATCKELASQPDVDGF   230
        ****      . : :   :                 :  *.. .         :    *

Escco   LVGGASLKADAFAVIVKAAEAAKQA----   255
Nossp   LVGGASLEAASFARIVNYL----------   241
Pyrfu   LLASGVTKAKDPEKAIWDLVSGIIKE---   228
Sulso   GVASAVMKAKEPEKVVEDFITSALRAISS   227
Sacce   LVGGASLK-PEFVDIINSRN---------   248
Arath   LVGGASLKGPEFATIVNSVTSKKVAA---   255
Homsa   LVGGASLK-PEFVDIINAKQ---------   249
         :...  :       :
```

Obr. 3.4. Zrovnanie vo formáte ALIGNMENT – „Clustal with character counts“, t.j. zrovnanie v blokoch s počítaním zvyškov, prevedené do textového editora (napr. MS-Word; t.j. súbor je uložený ako dokument). V tomto konkrétnom prípade – podľa zadania – napr. „TIM_aln.doc“. V súbore je možné robiť formálne úpravy, napr. farebné zvýraznenia funkčne dôležitých aminokyselinových zvyškov, konzervovaných sekvenčných regiónov, apod.

Konzenzuálna dĺžka

Na začiatku má každá sekvencia svoju vlastnú dĺžku vyjadrenú počtom aminokyselinových zvyškov. Pri zrovnávaní sa vyhľadávajú konzervované pozície – buď identické alebo aspoň podobné – ktoré sú usporiadané v zrovnaní pod sebou, t.j. dostanú sa na korešpondujúce pozície. Aby sa maximalizoval počet korešpondujúcich pozícií, do jednotlivých sekvencií sa zrovnávacím programom vkladajú tzv. medzery („gaps“), ktoré sú reprezentované v zrovnaných sekvenciách pomlčkami.

Konsenzuálna dĺžka („consensual length“; CL) je dĺžka sekvencií po ich zrovnaní. Minimálne sa rovná dĺžke najdlhšej sekvencie zo súboru sekvencií, ktoré sú zrovnávané, ale len v tom jedinom prípade, ak by sa do nej nevložila ani jedna medzera. Zvyčajne sa pri zrovnávaní sekvencií vložia medzery aj do najdlhšej sekvencie (najmä ak sekvencie nie sú veľmi podobné), preto je konsenzuálna dĺžka spravidla vždy väčšia ako je dĺžka najdlhšej sekvencie.

Pre vzorové zadanie 7 študovaných triózafosfátizomeráz (tab. 3.1) je hodnota konsenzuálnej dĺžky: CL = 269 (napr. obr. 3.2, resp. obr. 3.3).

Sekvenčná identita

Sekvenčná identita („sequence identity“; SI) je vyjadrená ako suma všetkých identických pozícií (program Clustal-Omega ich označuje symbolom „*“), vztiahnutá k hodnote konsenzuálnej dĺžky. Udáva sa v percentách.

$$SI = \frac{\Sigma\,(*)}{CL} \times 100\ (\%)$$

Pre vzorové zadanie 7 študovaných triózafosfátizomeráz (tab. 3.1) je hodnota sekvenčnej identity: SI = (20/269) × 100 = 7,44 %.

Sekvenčná podobnosť

Sekvenčná podobnosť („sequence similarity“; SS) je definovaná ako suma všetkých identických, aj podobných pozícií (program Clustal-Omega ich označuje symbolmi „*“, resp. „:“ a „.“), vztiahnutá k hodnote konsenzuálnej dĺžky. Udáva sa tiež v percentách.

$$SS = \frac{\Sigma (* + : + .)}{CL} \times 100\ (\%)$$

Pre vzorové zadanie 7 študovaných triózafosfátizomeráz (tab. 3.1) je hodnota sekvenčnej podobnosti: SS = [(20+27+16)/269] × 100 = (63/269) × 100 = 23,42 %.

Je potrebné si uvedomiť, že hodnota sekvenčnej identity je vždy menšia, maximálne rovná hodnote sekvenčnej podobnosti, pretože sekvenčná podobnosť zahŕňa aj podobné, aj identické aminokyselinové zvyšky v zrovnaní.

Z praktického hľadiska napr. hodnoty SI=7,44% a SS=23,42% pre vzorové zadanie 7 študovaných triózafosfátizomeráz (tab. 3.1) znamenajú, že na 100 pozícií v sekvenčnom zrovnaní je v priemere 92 pozícií, v ktorých sa toleruje zmena aminokyseliny pri súčasnom zachovaní funkcie enzýmu (triózafosfátizomerázy). Inými slovami, sekvenčnú identitu pod 10% možno považovať skôr za nízku, aj keď si je potrebné tiež uvedomiť, že funkčne dôležité zvyšky (ktorých však nebýva veľa – napr. pri enzýmoch sú to zvyšky katalytickej mašinérie, prípadne pár ďalších zvykov aktívneho miesta) ostanú pri takejto nízkej hodnote konzervované. Zároveň, aj pri nízkej hodnote sekvenčnej identity, môže byť hodnota sekvenčnej podobnosti relatívne vysoká; čo je aj prípad študovaných 7 triózafosfátizomeráz zo vzorového zadania, kde sa hodnota sekvenčnej podobnosti blíži k 25%.

Výpočet evolučných stromov

Pokiaľ ide o výpočty evolučných stromov, tieto sa počítajú na základe zrovnaných sekvencií. Existuje viacero prístupov, na ktorých sú výpočty evolučných vzťahov založené (napr. metódy UPGMA, Neighbour-joining, Maximum likelihood, Maximum parsimony a Minimum evolution). Pre účely získania prvotných skúseností s uskutočňovaním základnej bioinformatickej analýzy proteínov je dostačujúce oboznámiť sa prípravou evolučných stromov v rámci balíka programu Clustal na serveri EBI, t.j. v rámci programu Simple Phylogeny. Tento ponúka výpočet stromov dvomi

jednoduchšími metódami, a to klastrovacie metódy UPGMA („Unweighted Pair Group Method with Arithmetic Mean") a Neighbour-joining (metóda spájania susedov). Z nich bude používaná práve Nieghbour-joining metóda, ktorá je predvolenou metódou v programe Simple Phylogeny.

Vstupným súborom pre výpočet evolučných stromov je zrovnanie vo formáte Pearson/FASTA (obr. 3.3), z ktorého sa v dvoch postupných krokoch získajú dva evolučné stromy, resp. ich súbory (tab. 3.2):

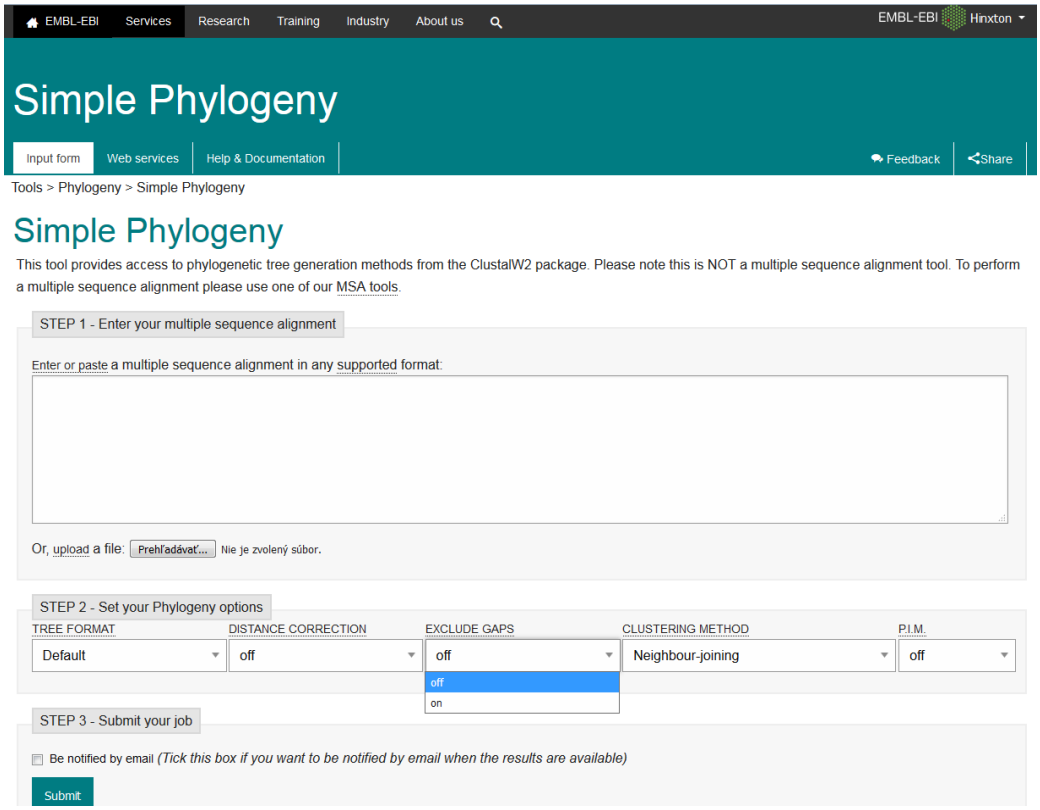
(i) súbor stromu založený na uvažovaní pozícií s medzerami v zrovnaných sekvenciách (súbor „TIM_off.txt)" – do výpočtu sa berú všetky pozície v zrovnaní;

(ii) súbor stromu založený na ignorovaní pozícií s medzerami v zrovnaných sekvenciách (súbor „TIM_on.txt)" – do výpočtu sa neberú pozície v zrovnaní, v ktorých sa nachádajú medzery.

V prvom prípade treba v programe zvoliť podmienku „Exclude gaps" ako „off" – t.j. treba ju vypnúť, kým v druhom prípade je treba túto podmienku zvoliť ako „on" – t.j. treba ju zapnúť.

Pre zrovnávané sekvencie, ktoré majú vysoký stupeň sekvenčnej podobnosti (identity), t.j. pri zrovnávaní bolo do nich vložených málo medzier, stromy počítané s uvažovaním pozícií s medzerami a s ich ignorovaním z toho istého zrovnania si budú veľmi podobné. Pokiaľ však sekvencie vykazujú nízky stupeň vzájomnej podobnosti, t.j. ich zrovnanie obsahuje veľa pozícií s medzerami, stromy môžu byť hodne odlišné. Ide o to, že ak sa v sekvenčnom zrovnaní nachádza príliš veľa pozícií s medzerami, pri výpočte evolučného stromu s ich ignorovaním môže slúžiť ako základ pre výpočet stromu len málo pozícií zrovnania v porovnaní s jeho celou, t.j. konsenzuálnou dĺžkou. To znamená, že pri výpočte evolučného stromu s uvažovaním pozícií s medzerami sa berú do úvahy aj podobnosti, aj rozdiely v sekvenciách, kým pri výpočte stromu založenom na ignorovaní pozícií s medzerami sa pozornosť sústreďuje skôr na to, čo dané sekvencie spája, t.j. čo majú spoločné (podobné) – čo je konzervované u všetkých sekvencií z daného študovaného súboru – za súčasného možného ignorovania väčšiny rozdielnych čŕt medzi sekvenciami. Treba pripomenúť, že obidva prístupy majú takto svoje výhody, aj nevýhody.

Internetové prostredie programu Simple Phylogeny na EBI serveri:
http://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/

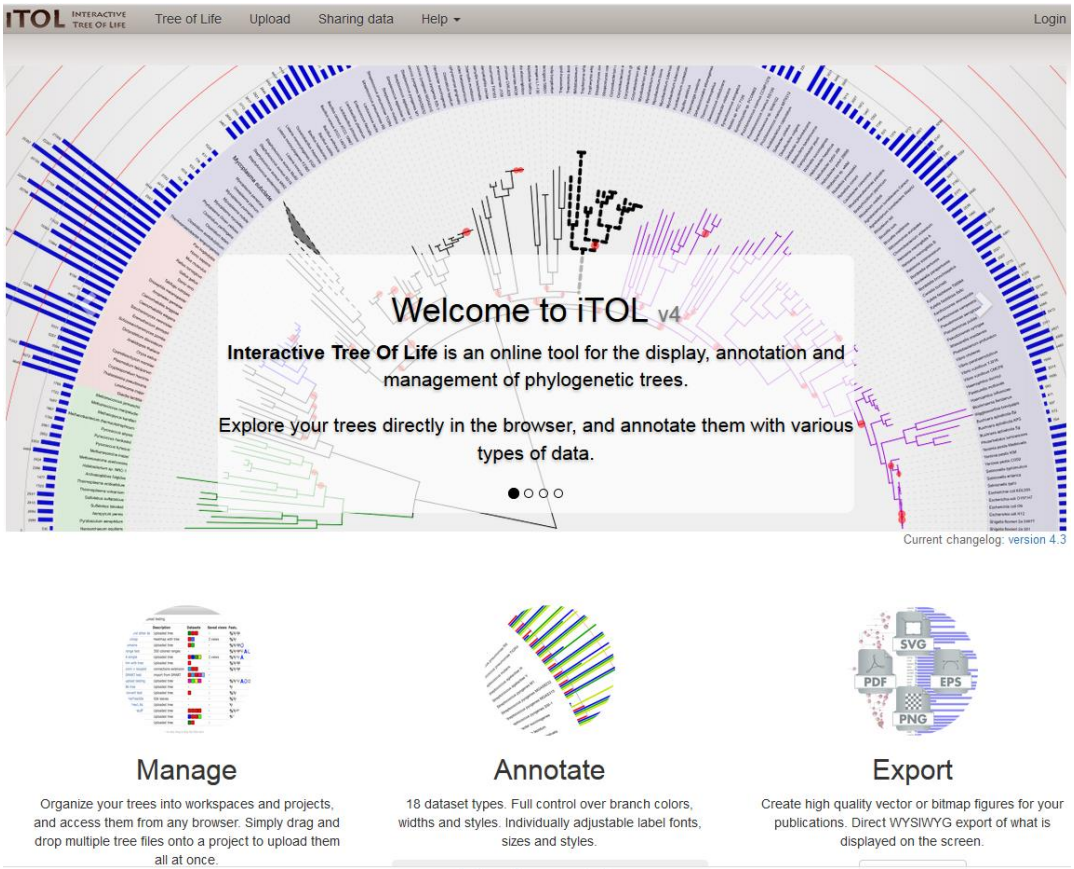


Tabuľka 3.2. Vypočítané súbory evolučných stromov.

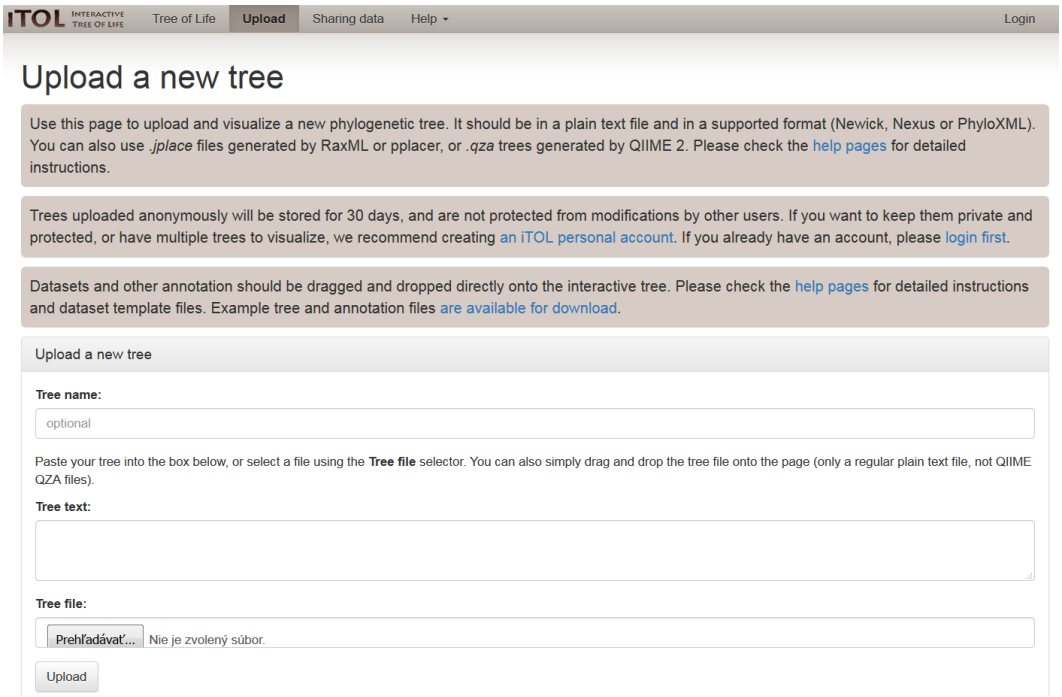
TIM_off.txt	TIM_on.txt
(Escco:0.26481, (Nossp:0.31843, (Pyrfu:0.30106, Sulso:0.27483) :0.19220) :0.02373, (Sacce:0.24481, (Arath:0.19143, Homsa:0.18760) :0.04340) :0.04189);	((Escco:0.25239, (Sacce:0.22289, (Arath:0.18242, Homsa:0.17644) :0.03309) :0.04785) :0.02512, Nossp:0.30981, (Pyrfu:0.29665, Sulso:0.26316) :0.21172);

Úvodná web-stránka programu iTOL (interactive Tree of Life)

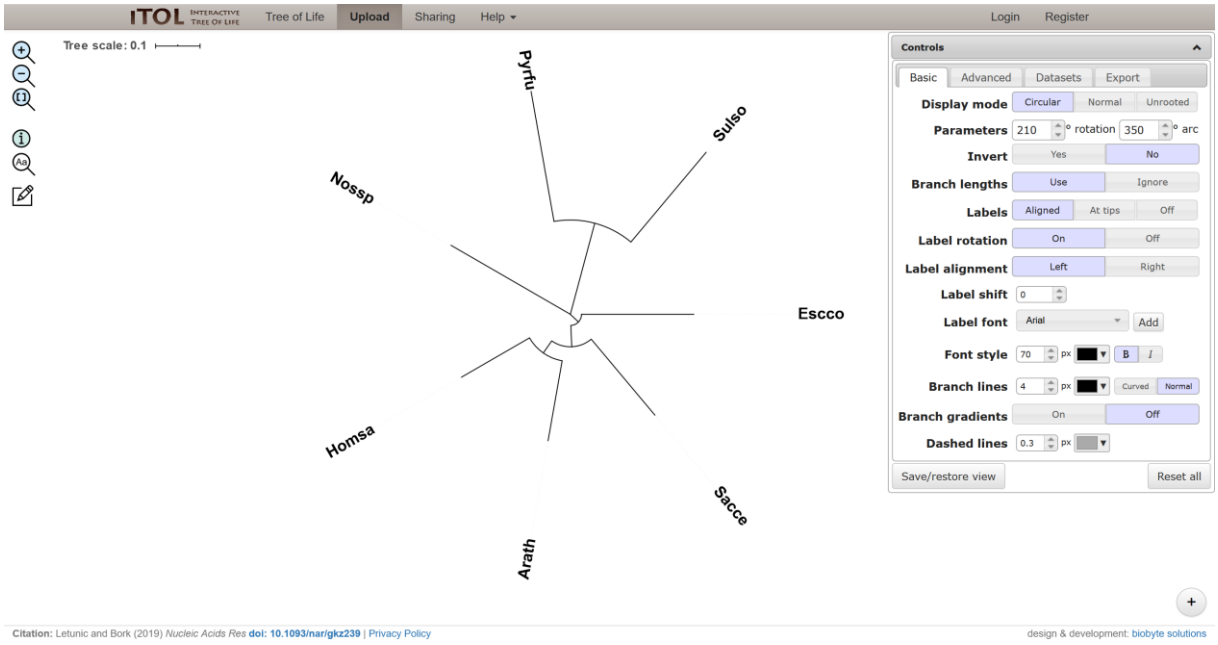
https://itol.embl.de/



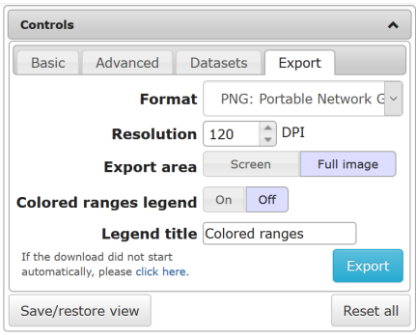
Prostredie programu iTOL pre načítanie súboru evolučného stromu:



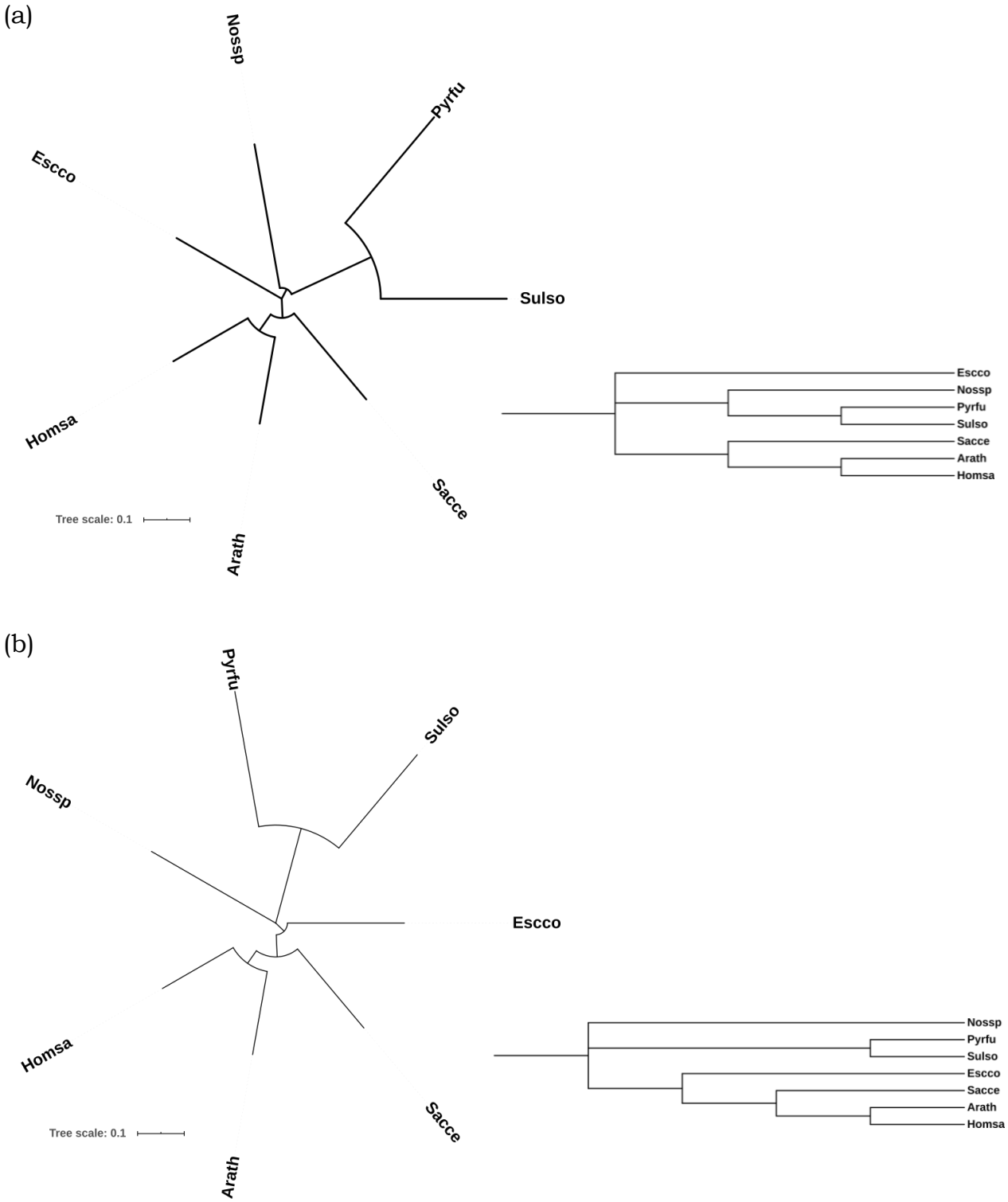
Prostredie programu iTOL – možnosti zobrazenia evolučného stromu:



Možnosti, pri exportovaní súboru obrázku evolučného stromu:



Pre vzorové zadanie 7 študovaných triózafosfátizomeráz (tab. 3.1) sú obidva evolučné stromy – počítané na základe zrovnania s uvažovaním pozícií s medzerami, aj s ich ignorovaním – znázornené na obr. 3.5. Oba stromy sú v podstate veľmi podobné. Na stromoch sú tri zoskupenia (klastre) triózafosfátizomeráz z týchto organizmov: (i) archeóny – *Pyrococcus furiosus* (Pyrfu) a *Sulfolobus solfataricus* (Sulso); (ii) eukaryoty – *Homo sapiens* (Homsa), *Arabidopsis thaliana* (Arath) a *Saccharomyces cerevisiae* (Sacce); a (iii) baktérie – *Escherichia coli* (Escco) a *Nostoc* sp. PCC 7120 (Nossp). Sekvencie eukaryotických triózafosfátizomeráz sú si evidentne viac podobné (príbuzné) ako dvojica archeálnych triózafosfátizomeráz, i keď obe dvojice tvoria na evolučných stromoch vzájomne najbližšie príbuzný pár sekvencií.

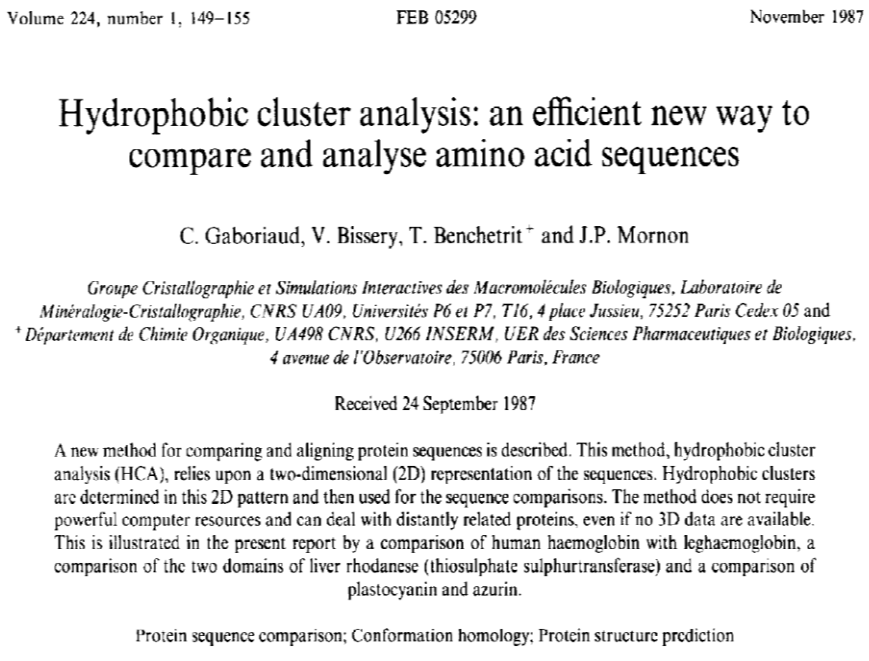


Obr. 3.5. Ilustrácie evolučných stromov. V ľavej časti obrázku sú tzv. kruhové („circular“) typy a v pravej časti sú tzv. pravouhlé („rectangular“) typy zobrazenia evolučných stromov. V prvom prípade je dôležitá aj samotná dĺžka vetiev (je tam udaná mierka, ktorej dĺžka udáva zmenu 0,1 aminokyseliny na dané miesto), kým v druhom prípade dĺžka vetiev je len pomerná (jej dĺžka nemá reálny význam). Zobrazené súbory sú konkrétne súbory pre vzorové zadanie 7 študovaných triózafosfátizomeráz, t.j. (a) TIM_off.txt“ a (b) „TIM_on.txt“.

4. HCA – metóda analýzy hydrofóbnych klastrov

Metóda analýzy hydrofóbnych klastrov – HCA („Hydrophobic Cluster Analysis“) – bola vyvinutá a predstavená ako nová metóda na porovnanie a zrovnávanie sekvencií proteínov v roku 1987 (obr. 4.1). Je to metóda použiteľná iba na aminokyselinové sekvencie. Jej podstatou je dvoj-rozmerná reprezentácia sekvencie proteínu, v ktorej obraze sú určené hydrofóbne klastre; pričom tieto dvoj-rozmerné vzory rozloženia hydrofóbnych aminokyselinových zvyškov slúžia ďalej na porovnanie sekvencií.

K výhodám metódy HCA patrí, že: (i) nevyžaduje k svojej realizácii výkonnú výpočtovú techniku; (ii) je vhodná aj na analýzu vzdialene príbuzných proteínov (t.j. pri vysokom stupni ich divergovanosti); (iii) je senzitívnejšia ako bežné zrovnávacie programy (napr. Clustal-Omega, apod.); a (iv) možno ju využiť aj pri absencii údajov o terciárnej štruktúre študovaných proteínov. Na druhej strane si však jej spoľahlivé využitie vyžaduje získať väčšiu rutinu a prax, ako pri práci s bežnými programami na zrovnávanie sekvencií. Na obr. 4.2 sú titulné stránky publikácií, v ktorých bola HCA metóda predstavená po niekoľkých rokoch úspešného používania.



Volume 224, number 1, 149–155 FEB 05299 November 1987

Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences

C. Gaboriaud, V. Bissery, T. Benchetrit + and J.P. Mornon

Groupe Cristallographie et Simulations Interactives des Macromolécules Biologiques, Laboratoire de Minéralogie-Cristallographie, CNRS UA09, Universités P6 et P7, T16, 4 place Jussieu, 75252 Paris Cedex 05 and *+ Département de Chimie Organique, UA498 CNRS, U266 INSERM, UER des Sciences Pharmaceutiques et Biologiques, 4 avenue de l'Observatoire, 75006 Paris, France*

Received 24 September 1987

A new method for comparing and aligning protein sequences is described. This method, hydrophobic cluster analysis (HCA), relies upon a two-dimensional (2D) representation of the sequences. Hydrophobic clusters are determined in this 2D pattern and then used for the sequence comparisons. The method does not require powerful computer resources and can deal with distantly related proteins, even if no 3D data are available. This is illustrated in the present report by a comparison of human haemoglobin with leghaemoglobin, a comparison of the two domains of liver rhodanese (thiosulphate sulphurtransferase) and a comparison of plastocyanin and azurin.

Protein sequence comparison; Conformation homology; Protein structure prediction

Obr. 4.1. Úvodná publikácia vo *FEBS Letters* z roku 1987 o metóde HCA.

Biochimie (1990) 72, 555–574
© Société française de biochimie et biologie moléculaire / Elsevier, Paris

Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences

L Lemesle-Varloot[1], B Henrissat[2], C Gaboriaud[1,3], V Bissery[1], A Morgat[1,4], JP Mornon*[1]

[1]*Laboratoire de Minéralogie-Cristallographie, Universités Paris 6 and 7, CNRS URA 09, T16, 4 place Jussieu, 75252 Paris Cedex 05;*
[2]*Centre de Recherches sur les Macromolécules Végétales, CNRS, Université Joseph Fourier, BP 53X, 38041 Grenoble;*
[3]*Institut de Chimie des Substances Naturelles, CNRS, 91190 Gif-sur-Yvette;*
[4]*Service de Modélisation Moléculaire, Centre de Recherches Rhône Poulenc Santé, 13 quai Jules Guesde, 94403 Vitry-sur-Seine Cedex, France*

(Received 28 June 1990; accepted 11 July 1990)

Summary – Hydrophobic cluster analysis (HCA) [15] is a very efficient method to analyse and compare protein sequences. Despite its effectiveness, this method is not widely used because it relies in part on the experience and training of the user. In this article, detailed guidelines as to the use of HCA are presented and include discussions on: the definition of the hydrophobic clusters and their relationships with secondary and tertiary structures; the length of the clusters; the amino acid classification used for HCA; the HCA plot programs; and the working strategies. Various procedures for the analysis of a single sequence are presented: structural segmentation, structural domains and secondary structure evaluation. Like most sequence analysis methods, HCA is more efficient when several homologous sequences are compared. Procedures for the detection and alignment of distantly related proteins by HCA are described through several published examples along with 2 previously unreported cases: the β-glucosidase from *Ruminococcus albus* is clearly related to the β-glucosidases from *Clostridium thermocellum* and *Hansenula anomala* although they display a reverse organization of their constitutive domains; the alignment of the sequence of human GTPase activating protein with that of the *Crk* oncogene is presented. Finally, the pertinence of HCA in the identification of important residues for structure / function as well as in the preparation of homology modelling is discussed.

protein sequences / protein structure / alignment / sequence comparisons / secondary structure / homology detection

CMLS, Cell. mol. life sci. 53 (1997) 621–645
1420-682X/97/080621-25 $ 1.50 + 0.20/0

CMLS Cellular and Molecular Life Sciences

Review

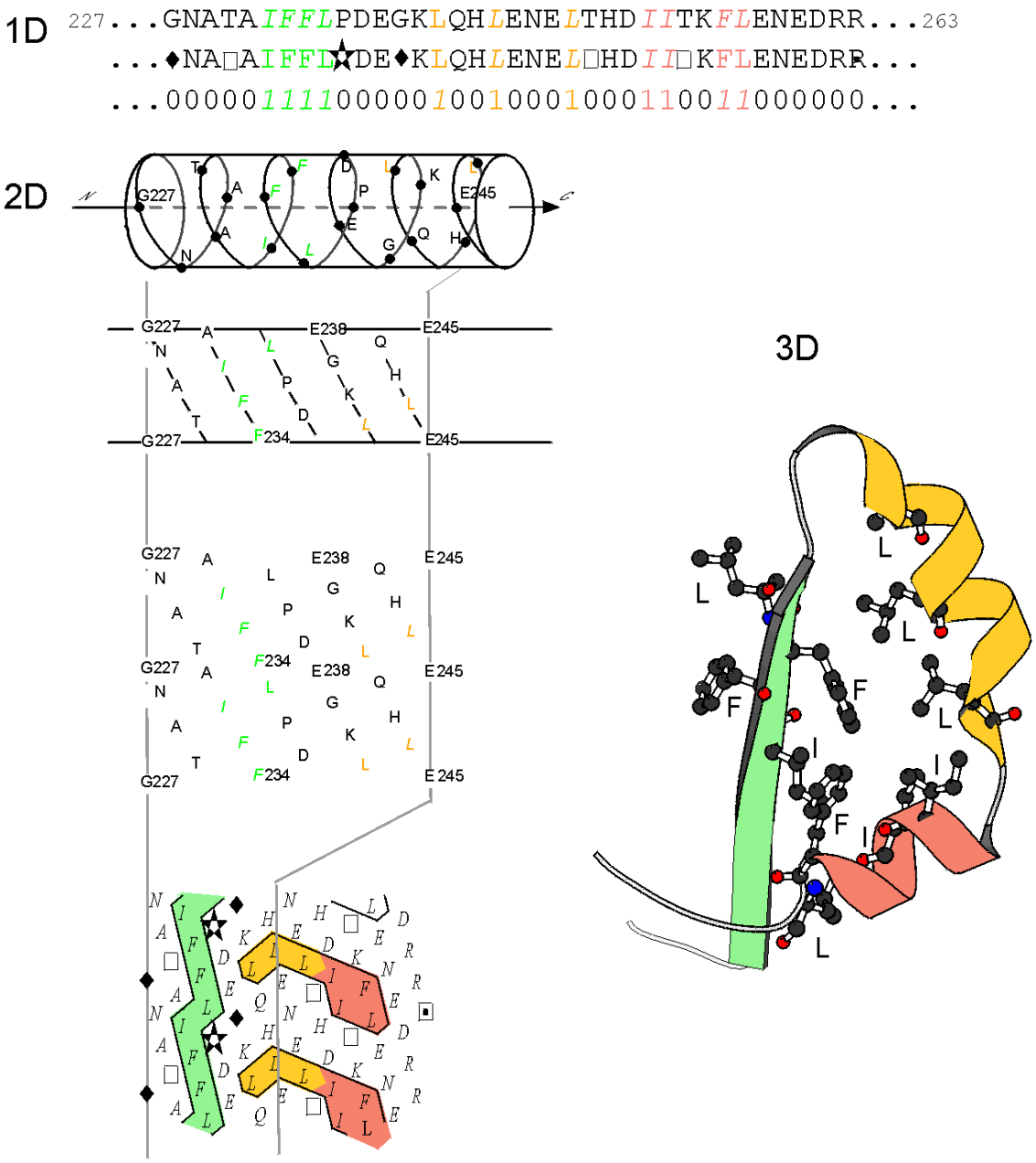
Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives

I. Callebaut[a], G. Labesse[a], P. Durand[a], A. Poupon[a], L. Canard[a], J. Chomilier[a], B. Henrissat[b] and J. P. Mornon[a,*]

[a]Systèmes Moléculaires et Biologie Structurale, LMCP, CNRS URA 09, UP6/UP7, Case 115, 4 place Jussieu, F-75252 Paris Cedex 05 (France), Fax +33 1 44 27 37 85, e-mail: Isabelle.Callebaut@lmcp.jussieu.fr
[b]Centre de Recherches sur les Macromolécules Végétales**, CNRS, BP53, F-38041 Grenoble Cedex 9 (France)

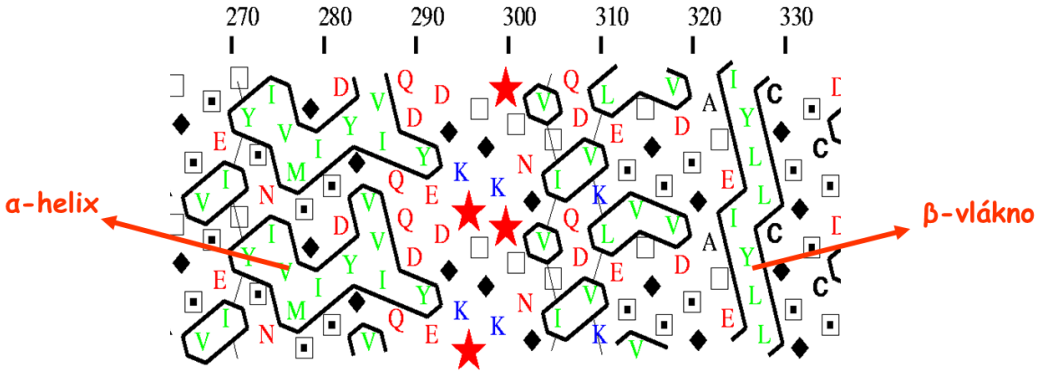
Obr. 4.2. Titulné stránky publikácií v časopisoch *Biochimie* a *Cell. Mol. Life Sci.* z rokov 1990, resp. 1997, v ktorých boli sumarizované pokroky v používaní metódy HCA.

Teoretické základy a postup jednotlivých krokov vedúcich k získaniu HCA obrazu sekvencie proteínu je vysvetlený na obr. 4.3.



Obr. 4.3. Popis jednotlivých krokov vedúcich k získaniu HCA obrazu sekvencie proteínu. Všeobecná aminokyselinová sekvencia (lineárna; 1D) je znázornená s farebne zvýraznenými hydrofóbnymi zvyškami. Vybraným aminokyselinám (Gly, Pro, Thr a Ser) sú priradené špeciálne symboly. Všetky aminokyselinové zvyšky sú rozdelené na hydrofóbne (hodnota „1“) a hydrofilné (hodnota „0“). Sekvencia je následne zapísaná (navinutá) na špirálu a zobrazená pozdĺž valca, ktorý je potom prerezaný paralelne so svojou osou (2D). Takto vzniknutý dvojrozmerný diagram je duplikovaný v snahe obnoviť celkové okolité prostredie každej aminokyseliny. Hydrofóbne zvyšky nie sú distribuované náhodne, ale tvoria klastre, ktoré sú zvýraznené programom pomocou obkreslenia. Pozície hydrofóbnych klastrov môžu korešpondovať s pozíciami pravidelných elementov sekundárnej štruktúry (α-helixy a β-vlákna), čo je znázornené na korešpondujúcej experimentálnej terciárnej štruktúre (3D). Upravené podľa Callebaut et al. (1997).

Okrem identifikácie prípadných korešpondencií medzi sekvenciami vzdialene príbuzných proteínov, možno metódu HCA – ale len v určitom zjednodušenom priblížení – využiť aj na čiastočnú predikciu sekundárnej štruktúry proteínov. Je to v dôsledku toho, že tvary hydrofóbnych klastrov môže korešpondovať s pravidelnými elementami sekundárnej štruktúry, ako sú α-helixy a β-vlákna. Horizontálne pretiahnutý tvar hydrofóbneho klastra môže indikovať prítomnosť α-helixu, kým vertikálne pretiahnutý tvar klastra môže byť spojený s existenciou β-vlákna v štruktúre proteínu (obr. 4.4).



Obr. 4.4. Ilustrácia horizontálneho a vertikálneho hydrofóbneho klastra v HCA obraze sekvencie proteínu, ktorá môže naznačovať α-helix, resp. β-vlákno v jeho štruktúre.

Pre lepšiu vizualizáciu a uľahčenie identifikácie korešpondencií medzi analyzovanými sekvenciami proteínov, štyri vybrané aminokyseliny majú špeciálne symboly: glycín – ◆; prolín – ★; treonín – □; a serín – ⊡.

V metóde HCA je 20 aminokyselín rozdelených do dvoch hlavných tried: (i) hydrofóbne; a (ii) hydrofilné a/alebo indiferentné k ich prostrediu. Toto rozdelenie je urobené arbitrážne, kvôli zjednodušeniu a možnosti analyzovať potenciálne hydrofóbne klastre, aj keď nie celkom odpovedá realite. V skutočnosti má každá aminokyselina svoju vlastnú hodnotu hydrofóbnosti, resp. hydrofilnosti, a to aj v závislosti od podmienok merania.

Pre globulárne proteíny platí, že prevažne hydrofóbne aminokyselinové zvyšky tvoria vnútorné jadro proteínu, kým prevažne hydrofilné zvyšky tvoria povrch (ochrana jadra pred vodou a iónmi). Toto rozdelenie je dôsledkom entropických a entalpických síl, ktoré ženú proces k tvorbe hydrofóbnych

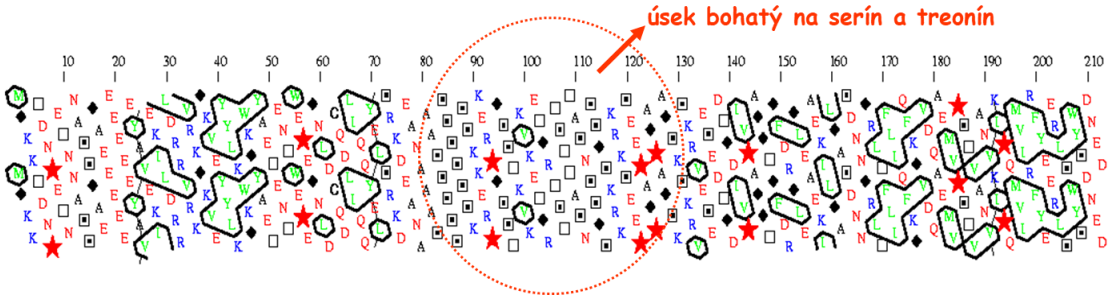
klastrov (interakcií) vo vnútorných oblastiach proteínu. Hydrofóbne interakcie sú jedným zo základov udržiavania terciárnej štruktúry proteínu. Hydrofóbne aminokyseliny sú uprednostňované vo vnútorných častiach pravidelných elementov sekundárnej štruktúry (α-helix a β-list), pričom sa menej vyskytujú v nepravidelných elementoch sekundárnej štruktúry (slučky a ohyby).

Na základe pozorovaní, skúseností a výpočtov bolo 20 aminokyselín rozdelených v metóde HCA nasledovne:

(i) silne hydrofóbne aminokyseliny (Val, Ile, Leu a Phe) – s hodnotou HCA: „1“ – ktoré sú hnacou silou tvorby hydrofóbnych, vnútorných častí elementov sekundárnej štruktúry;

(ii) stredne hydrofóbne aminokyseliny (Met, Trp a Tyr) – rovnako s hodnotou HCA: „1“ – každá so svojimi špecifickými vlastnosťami;

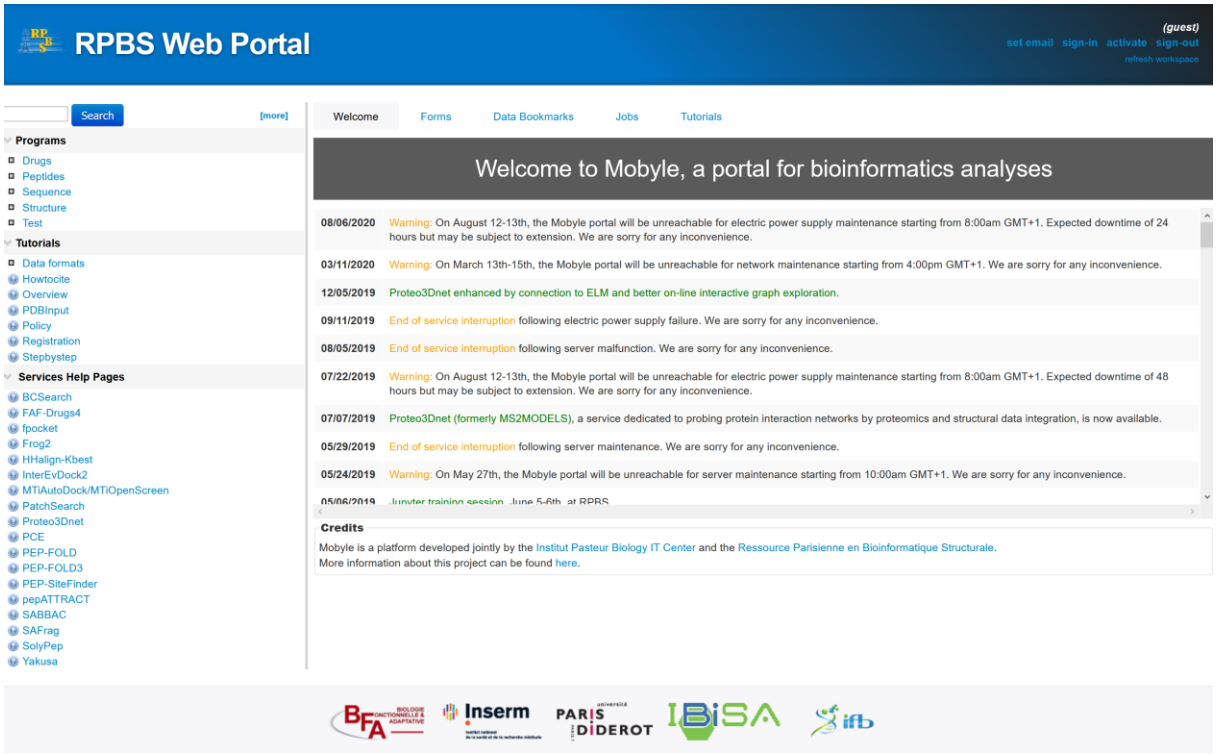
(iii) všetky ostatné aminokyseliny (Gly, Ala, Ser, Thr, Cys, His, Pro, Asp, Glu, Asn, Gln, Lys a Arg) – s hodnotou HCA: „0“ – sú považované za hydrofilné, resp. nie hydrofóbne.

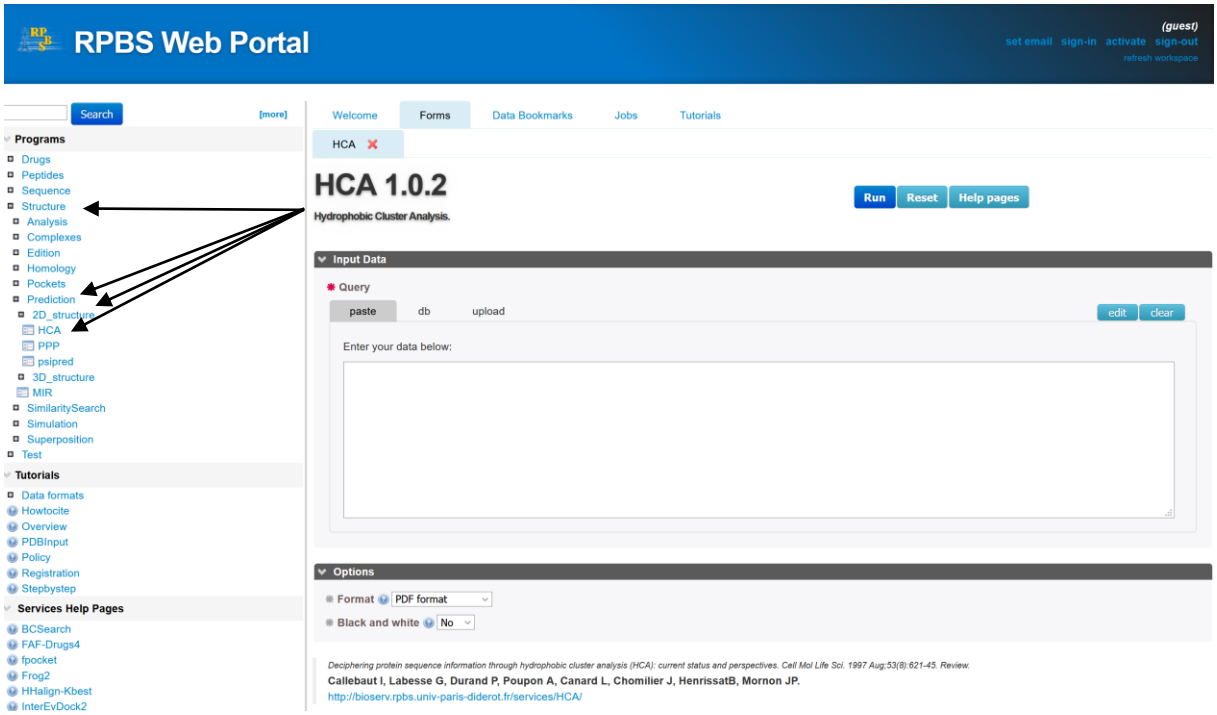


Obr. 4.5. Predikcia medzidoménovej oblasti proteínu – úsek bohatý na treonín a serín.

Aminokyselina prolín, ktorá často prerušuje sekundárne štruktúry, je považovaná za zvyšok prerušujúci hydrofóbne klastre. Cysteín neprerušuje klastrové správanie aminokyselín v proteínoch, ale tvorí S-S mostíky. Zvyšky treonín a serín často maskujú svoju polaritu a ich zvýšená prítomnosť môže signalizovať hraničné oblasti domén v proteínoch (obr. 4.5). Kyselina asparágová a kyselina glutámová sú na opačnej strane spektra ako valín, leucín, izoleucín a fenylalanín.

Metóda HCA – príprava HCA obrazu sekvencie proteínu – je dostupná na internete: https://mobyle.rpbs.univ-paris-diderot.fr/; v rámci „Programs“ postupom cez „Structure“ – „Prediction“ – „2D Structure“ – „HCA“.



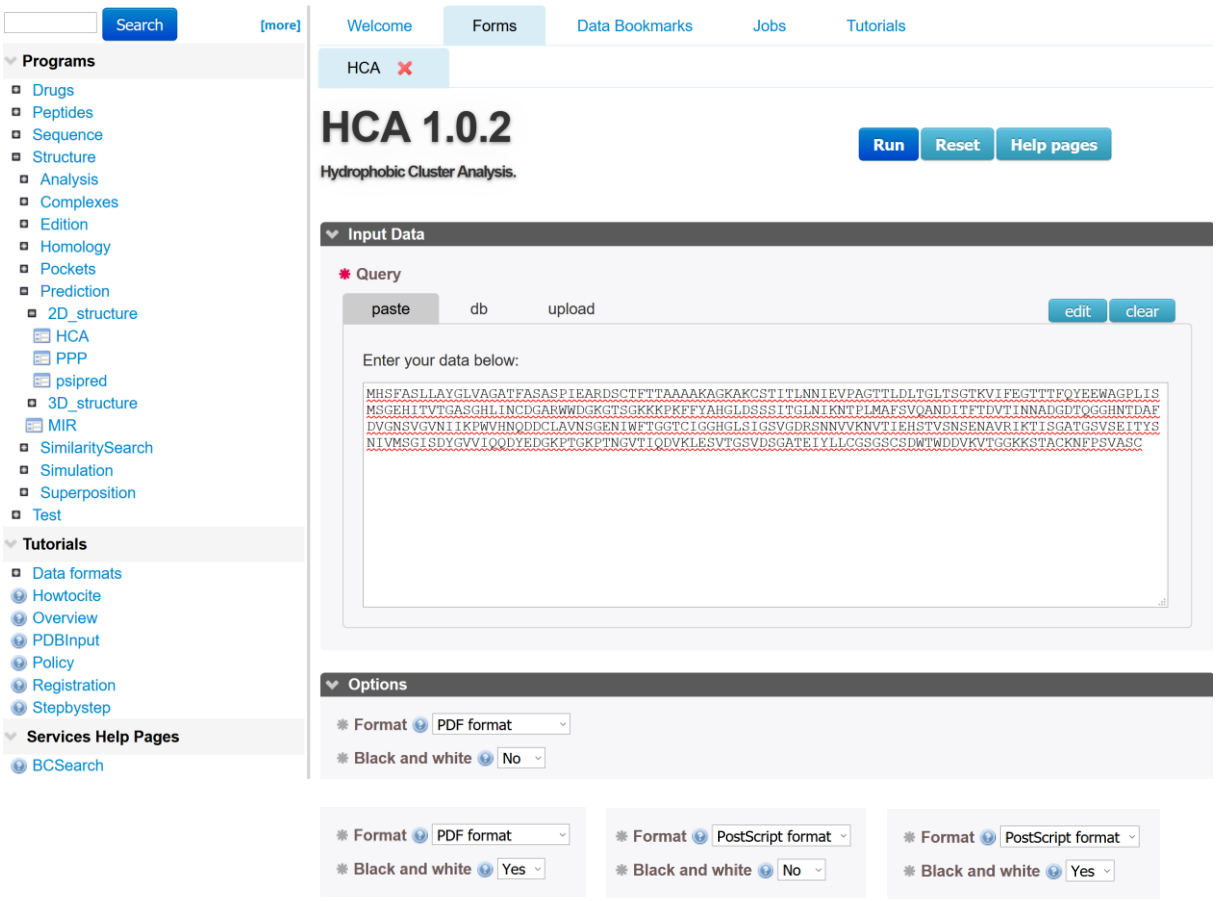


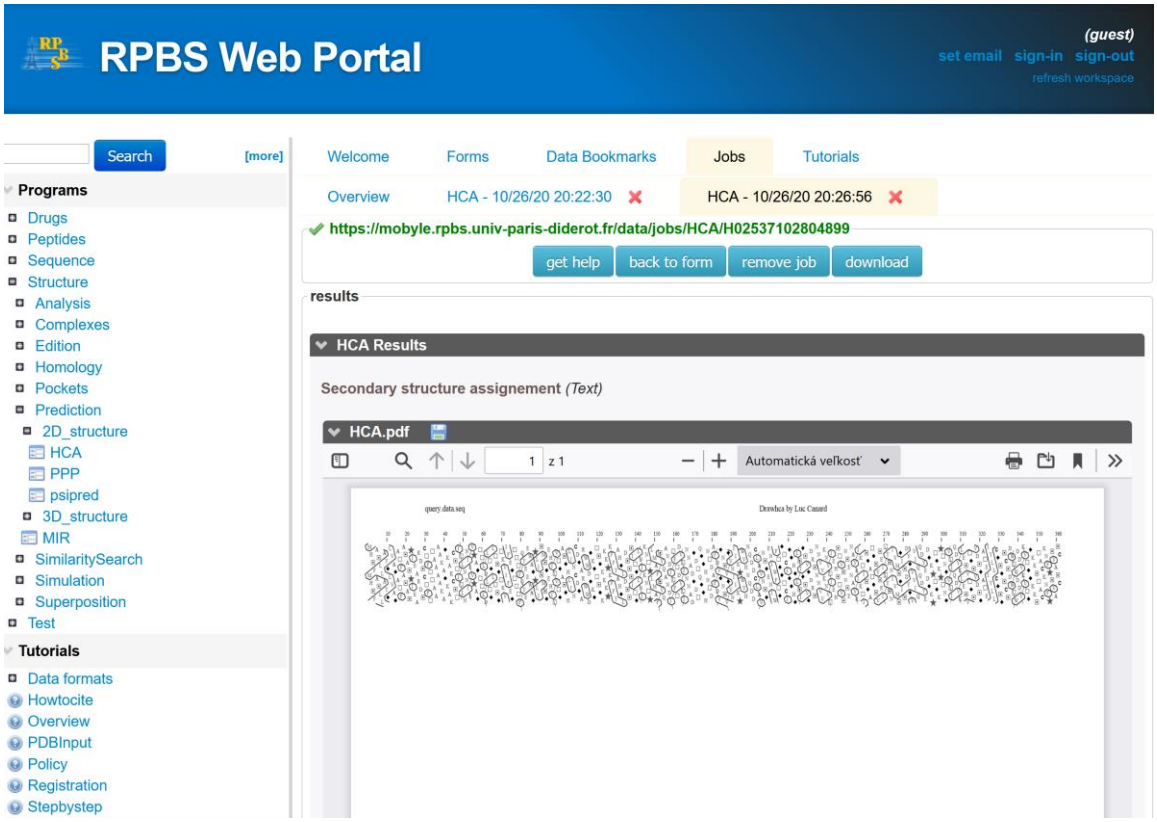
Nasledujúci príklad ilustruje prípravu HCA obrazu sekvencie polygalakturonázy z *Aspergillus niger*, sekvencia ktorej je uložená v databáze UniProt pod prístupovým číslom: P26214.

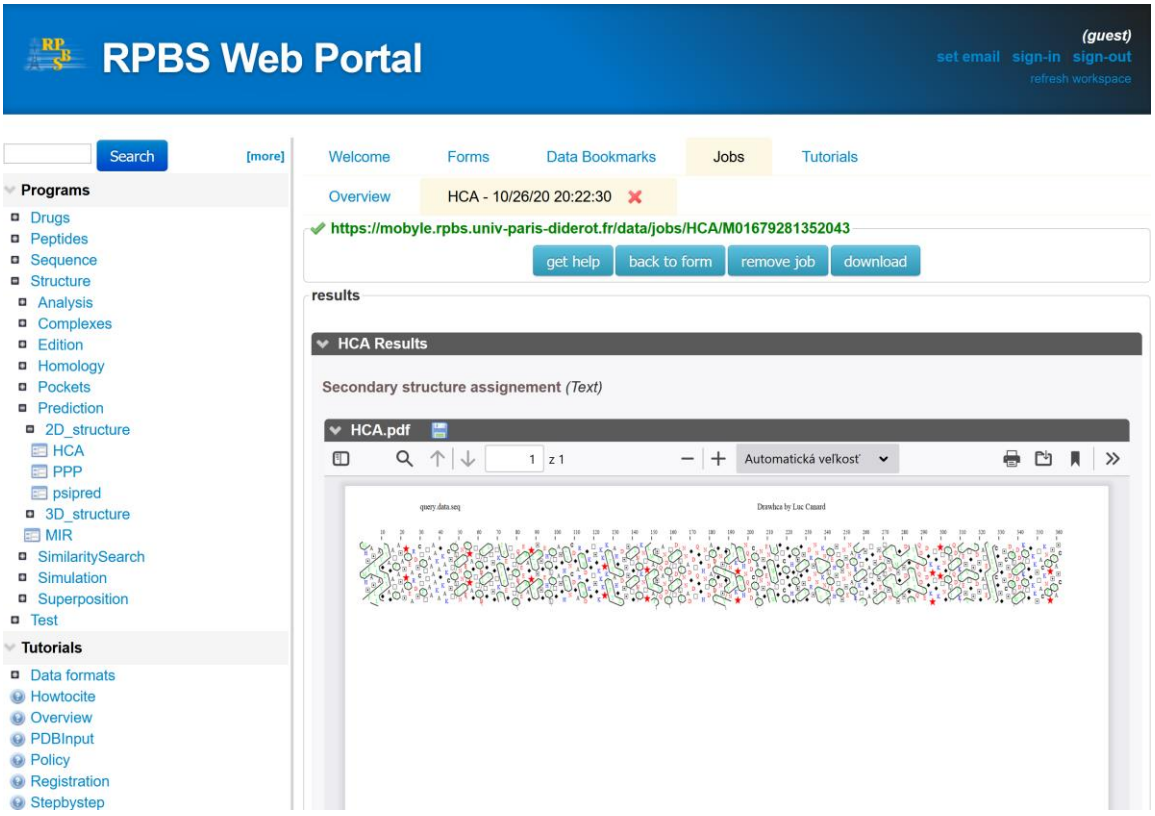
>sp|P26214|PGLR2_ASPNG Endopolygalacturonase-2 OS=Aspergillus niger

MHSFASLLAYGLVAGATFASASPIEARDSCTFTTAAAAKAGKAKCSTITLNNIEVPAGTT
LDLTGLTSGTKVIFEGTTTFQYEEWAGPLISMSGEHITVTGASGHLINCDGARWWDGKGT
SGKKKPKFFYAHGLDSSSITGLNIKNTPLMAFSVQANDITFTDVTINNADGDTQGGHNTD
AFDVGNSVGVNIIKPWVHNQDDCLAVNSGENIWFTGGTCIGGHGLSIGSVGDRSNNVVKN
VTIEHSTVSNSENAVRIKTISGATGSVSEITYSNIVMSGISDYGVVIQQDYEDGKPTGKP
TNGVTIQDVKLESVTGSVDSGATEIYLLCGSGSCSDWTWDDVKVTGGKKSTACKNFPSVA
SC

Sekvenciu – ako text – je potrebné vložiť ako vstupný údaj („query“) a výsledky, t.j. HCA obraz sekvencie proteínu, je možné získať ako „PDF“ alebo „PostScript“ súbor, a to buď v čierno-bielej verzii alebo farebom prevedení.




RPBS Web Portal
(guest)
set email sign-in sign-out
refresh workspace
Search
[more]
Welcome
Forms
Data Bookmarks
Jobs
Tutorials
Overview
HCA - 10/26/20 20:22:30
HCA - 10/26/20 20:26:56
https://mobyle.rpbs.univ-paris-diderot.fr/data/jobs/HCA/H02537102804899
get help
back to form
remove job
download
results
HCA Results
Secondary structure assignement (Text)
HCA.pdf
Programs
Drugs
Peptides
Sequence
Structure
Analysis
Complexes
Edition
Homology
Pockets
Prediction
2D_structure
HCA
PPP
psipred
3D_structure
MIR
SimilaritySearch
Simulation
Superposition
Test
Tutorials
Data formats
Howtocite
Overview
PDBInput
Policy
Registration
Stepbystep

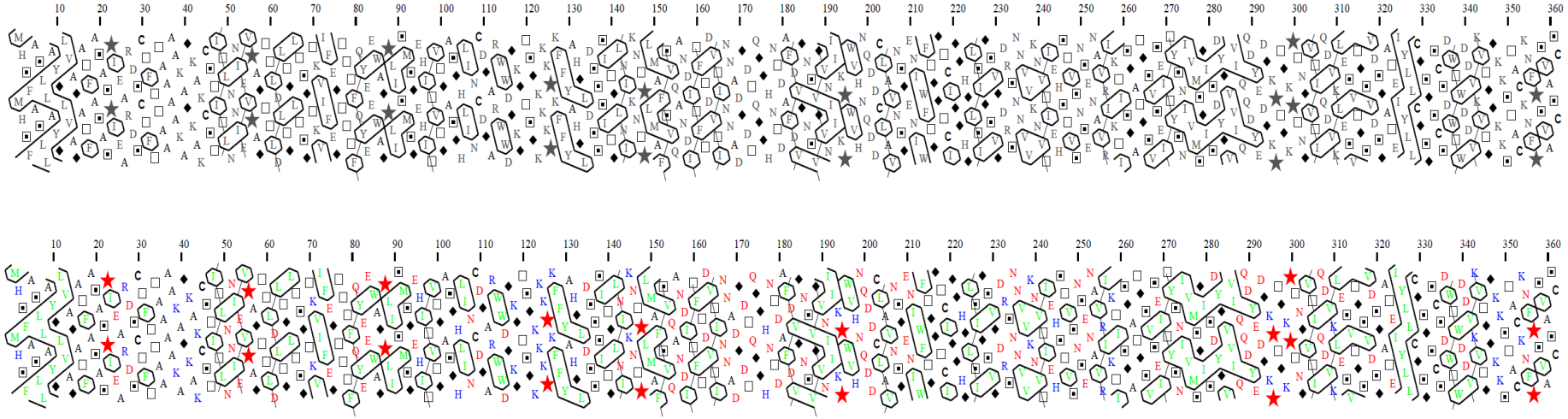

RPBS Web Portal
(guest)
set email sign-in sign-out
refresh workspace
Search
[more]
Welcome
Forms
Data Bookmarks
Jobs
Tutorials
Overview
HCA - 10/26/20 20:22:30
https://mobyle.rpbs.univ-paris-diderot.fr/data/jobs/HCA/M01679281352043
get help
back to form
remove job
download
results
HCA Results
Secondary structure assignement (Text)
HCA.pdf
Programs
Drugs
Peptides
Sequence
Structure
Analysis
Complexes
Edition
Homology
Pockets
Prediction
2D_structure
HCA
PPP
psipred
3D_structure
MIR
SimilaritySearch
Simulation
Superposition
Test
Tutorials
Data formats
Howtocite
Overview
PDBInput
Policy
Registration
Stepbystep

Vstupná sekvencia:

>sp|P26214|PGLR2_ASPNG Endopolygalacturonase-2 OS=Aspergillus niger

MHSFASLLAYGLVAGATFASASPIEARDSCTFTTAAAAKAGKAKCSTITLNNIEVPAGTT
LDLTGLTSGTKVIFEGTTTFQYEEWAGPLISMSGEHITVTGASGHLINCDGARWWDGKGT
SGKKKPKFFYAHGLDSSSITGLNIKNTPLMAFSVQANDITFTDVTINNADGDTQGGHNTD
AFDVGNSVGVNIIKPWVHNQDDCLAVNSGENIWFTGGTCIGGHGLSIGSVGDRSNNVVKN
VTIEHSTVSNSENAVRIKTISGATGSVSEITYSNIVMSGISDYGVVIQQDYEDGKPTGKP
TNGVTIQDVKLESVTGSVDSGATEIYLLCGSGSCSDWTWDDVKVTGGKKSTACKNFPSVA
SC

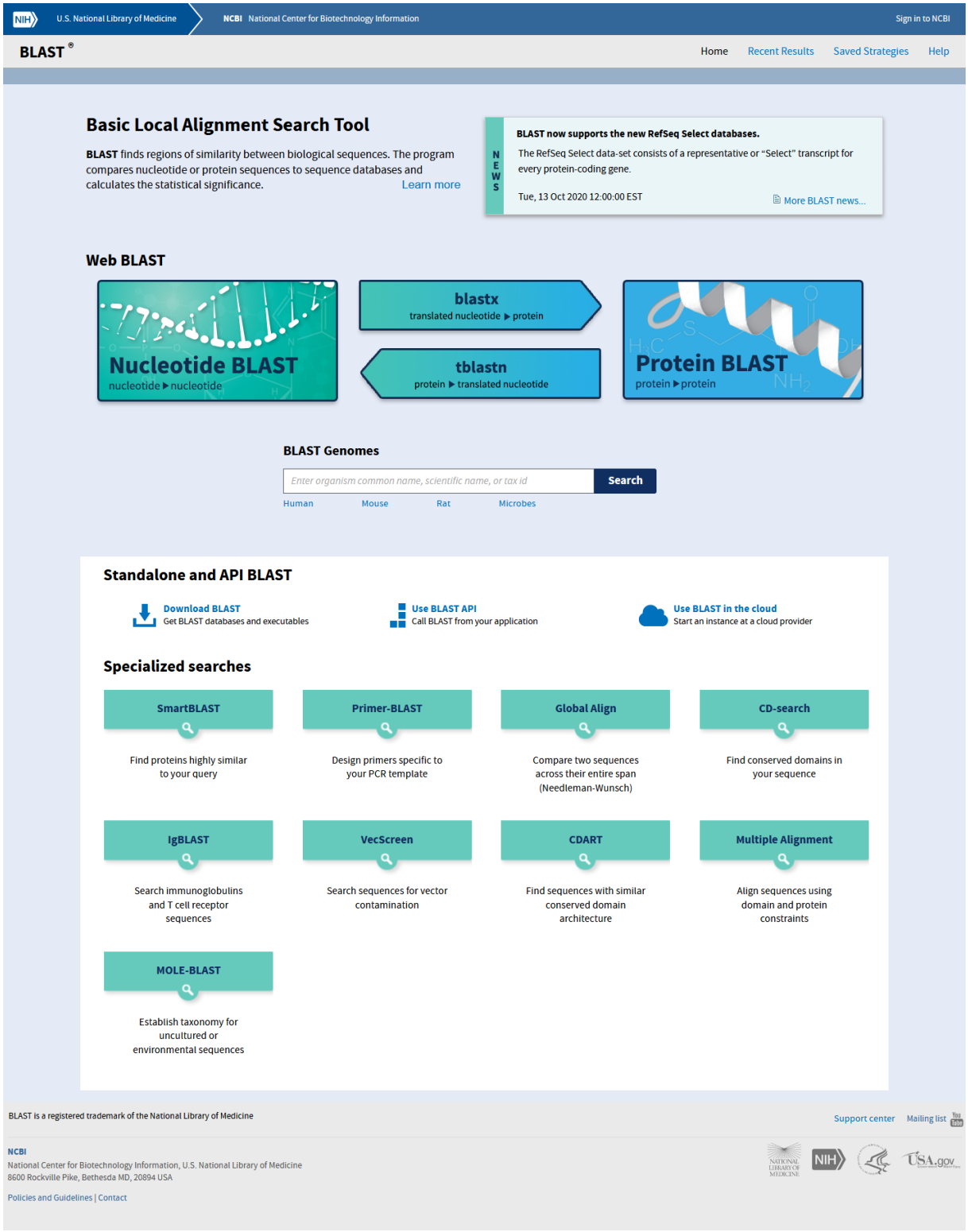
Výsledný HCA obraz sekvencie v čierno-bielej a farebnej verzii:



5. BLAST – nástroj na vyhľadávanie sekvenčných podobností

Úvodná web-stránka nástroja BLAST:

https://blast.ncbi.nlm.nih.gov/



BLAST („Basic Local Alignment Search Tool“) je internetový nástroj, ktorý hľadá regióny lokálnej podobnosti medzi sekvenciami. Program porovnáva nukleotidové alebo aminokyselinové sekvencie k sekvenčným databázam a počíta štatistickú významnosť zhody. BLAST možno použiť na odvodenie funkčných a evolučných vzájomných vzťahov medzi sekvenciami, ako aj ako pomôcku pri identifikovaní génových a proteínových rodín. BLAST bol originálne vyvinutý a predstavený vedeckej komunite v roku 1990; jeho domovská web-stránka je v rámci systému Entrez na serveri NCBI.

K základnému programovému vybaveniu patria štandardné typy BLASTn a BLASTp, ktoré prezerajú nukleotidové databázy s použitím záujmovej sekvencie („query“) nukleotidovej, resp. proteínové databázy s použitím záujmovej sekvencie („query“) aminokyselinovej. Ako *query* je označovaná sekvencia, o BLAST ktorej je záujem, t.j. možno ju považovať aj za dopytovu sekvenciu, prípadne žiadanku. K ďalším základným programom patria aj programy BLASTx a tBLASTn, ktoré prehľadávajú proteínové databázy použijúc preložený nukleotidový dopyt, resp. preložené nukleotidové databázy použijúc aminokyselinový dopyt. Prostredie BLAST ponúka aj rôzne špecializované vyhľadávania.

Proteínový BLAST

Samotný proteínový BLAST možno realizovať okrem základného štandardného algoritmu aj ako tzv. PSI-BLAST („Position-Specific Iterated“), PHI-BLAST („Pattern Hit Initiated“) a DELTA-BLAST („Domain Enhanced Lookup Time Accelerated“).

BLAST ponúka veľké možnosti, ako zamerať prehľadávanie databáz so známymi sekvenciami. Je možné blastovať query, ktorá predstavuje celú sekvenciu alebo len jej časť. Tiež je v ponuke výber databáz – ako prednastavená je databáza *Non-redundant protein sequences* (označená ako „nr“), ktorej voľba by mala zabezpečiť, že získané výsledky nebudú zbytočne redundantné, t.j. nadbytočné (duplicitné) v dôsledku viacerých záznamov pre tú istú sekvenciu, napr. aj pre ten istý záznam v databázach GenBank a UniProt. Ďalšou skvelou možnosťou je ponuka zamerať prehľadávanie len

na vybraný taxón, resp. taxóny (prípadne aj na celé ríše *Bacteria*, *Archaea* alebo *Eucarya*), ako aj možnosť ich vylúčenia z prehľadávania.

Dôležitým parametrom je „Max target sequences“, ktorým je možné obmedziť počet sekvencií, ktoré budú zobrazené vo výsledkoch; ich hodnotu je možné nastaviť od 10 do 5 000 (v nedávnej dobe to bolo ešte 20 tisíc).

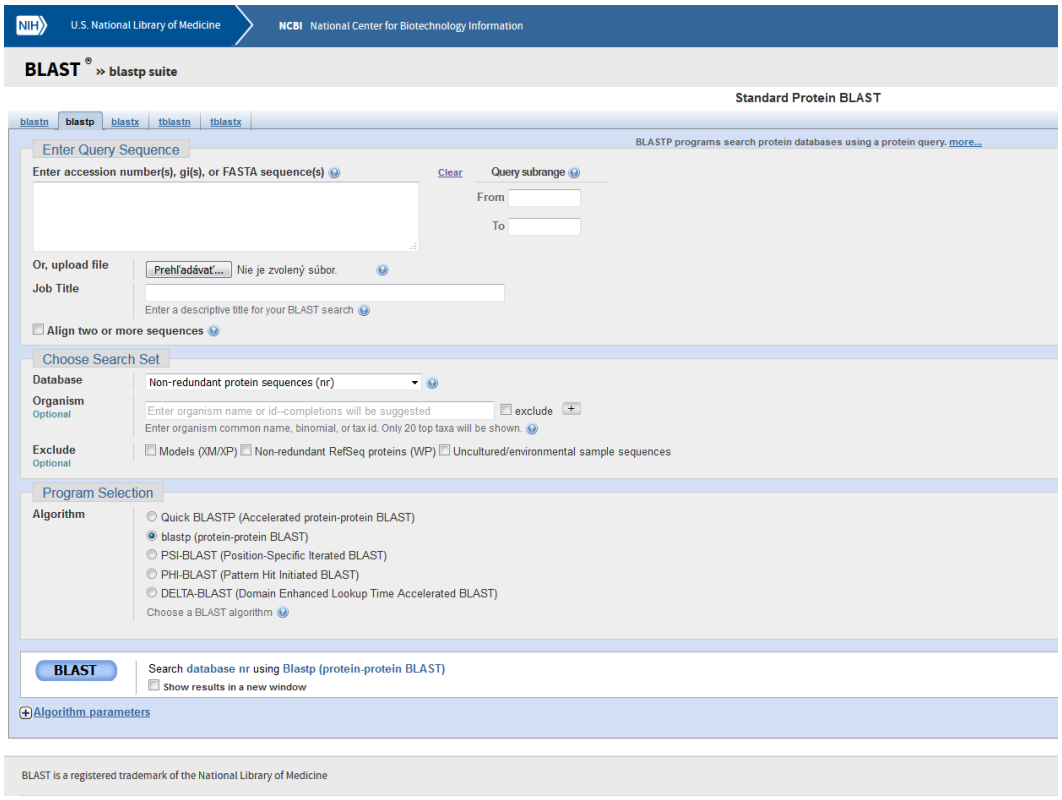
Pri opakovanom blastovaní s tou istou sekvenciou, ale pri menení podmienok a parametrov, je veľmi výhodnou funkciou možnosť „Show results in a new window“, t.j. možnosť zobraziť výsledky v novom okne monitoru.

Výsledky, ktoré poskytuje BLAST, sú štandardne zložené z troch, resp. štyroch častí: (i) opisy („descriptions“); (ii) grafický súhrn („graphic summary“); (iii) zrovnania („alignments“); a v súčasnosti aj (iv) taxonómia („taxonomy“). Za jeden z najdôležitejších výsledkov BLASTu možno považovať potenciálnu identifikáciu konzervovaných domén v sekvencii blastovaného proteínu, najmä, ak je málo informácií o jeho možnej funkcii.

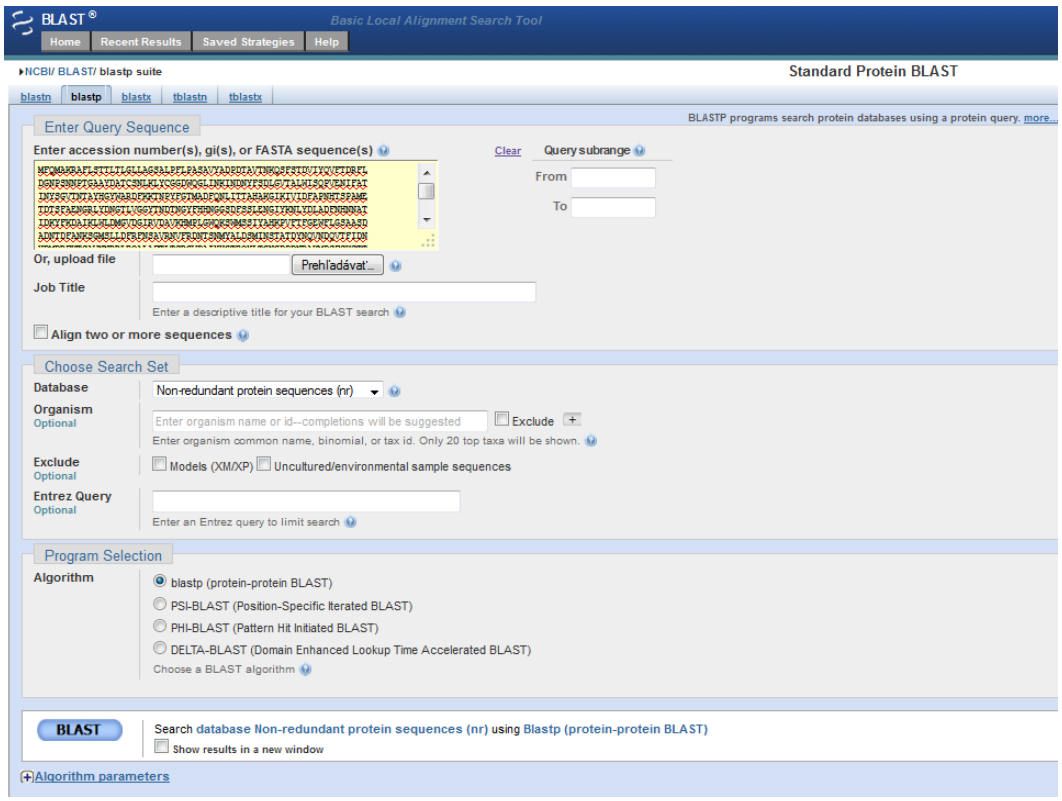
Štatistická významnosť zhody sa udáva ako skóre zrovnania – hodnota S („alignment score“), ktorého vyššia hodnota znamená vyššiu podobnosť medzi sekvenciami. Ďalšou hodnotou charakterizujúcou významnosť zásahu pri blastovaní je hodnota E („expectation“), ktorá predstavuje počet rôznych zrovnaní so skóre ekvivalentným alebo lepším ako to, pri ktorom sa očakáva, že sa pri vyhľadávaní v databáze vyskytne náhodne. V prípade hodnoty E platí, že čím nižšia je jeho hodnota, tým vyššia (významnejšia) je hodnota S, a teda aj samotného zrovnania. Pre najvýznamnejšie zásahy sa E hodnota limitne blíži k nule, naopak môže dosahovať hodnoty až v jednotkách.

V samotných zrovaniach BLASTu sa dopytová sekvencia (query) nachádza v 1. riadku, kým zachytená sekvencia (označovaná ako „subject“ – skr. „sbjct“) sa nachádza v 3. riadku. V strednom riadku medzi *query* a *subject* sú zvýraznené identické („identities“) a podobné („positives“) aminokyselinové zvyšky, pričom prvé sú udávané písmenom danej identickej aminokyseliny a druhé znamienkom plus (+). Dôležité je uvedomiť si aj to, že tzv. prvý zásah, („the first hit“) je spravidla samotná query, t.j. blastovaná sekvencia (ak už pochádza z databáz); pokiaľ sa teda neblastuje s unikátnou (novou) sekvenciou, ktorá sa v databázach ešte nenachádza.

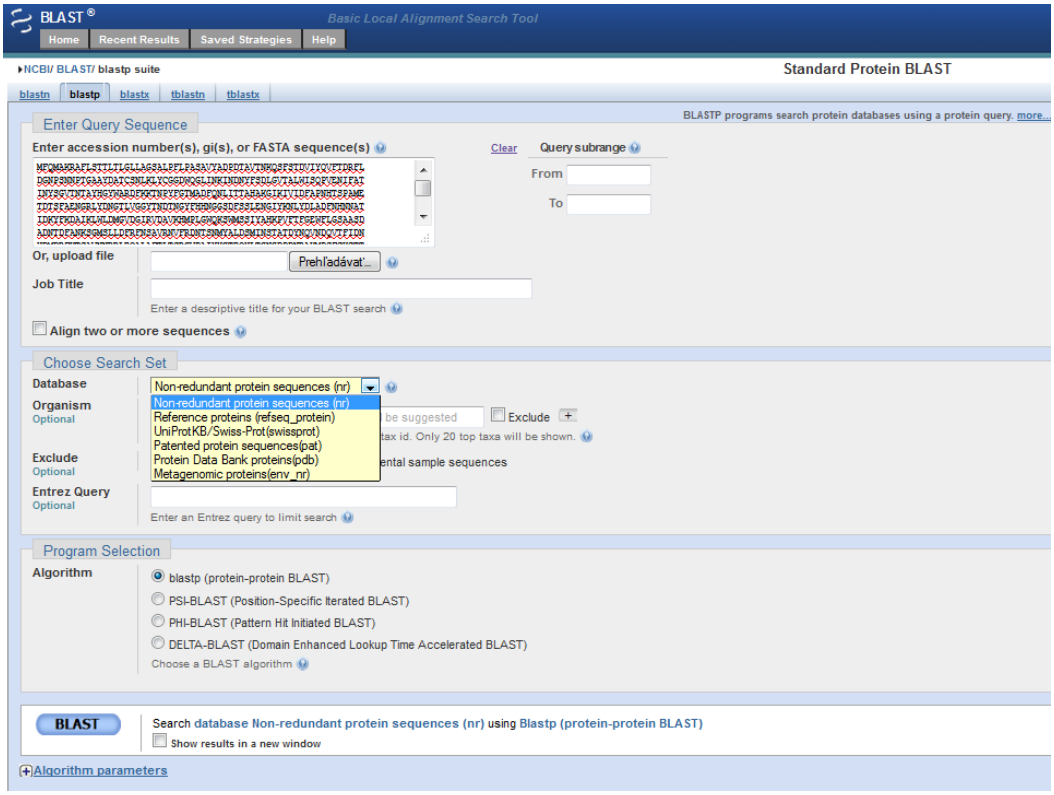
Úvodná web-stránka pre proteínový BLAST:



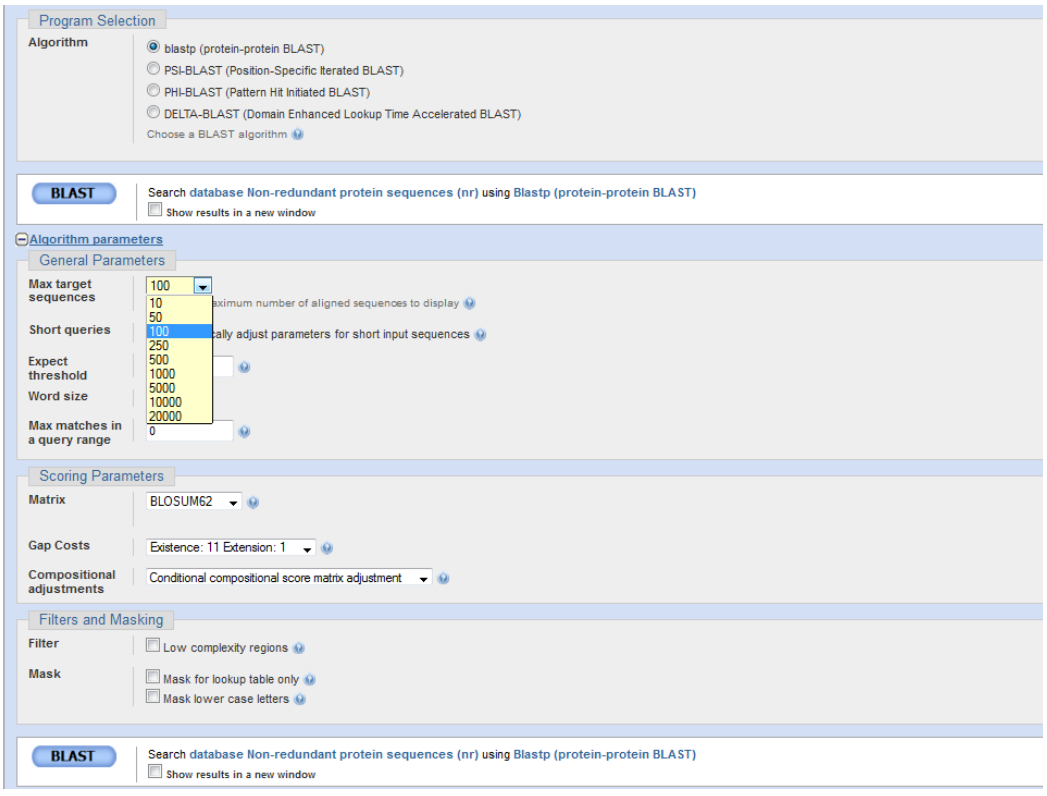
- načítanie sekvencie – query:



- výber databázy:

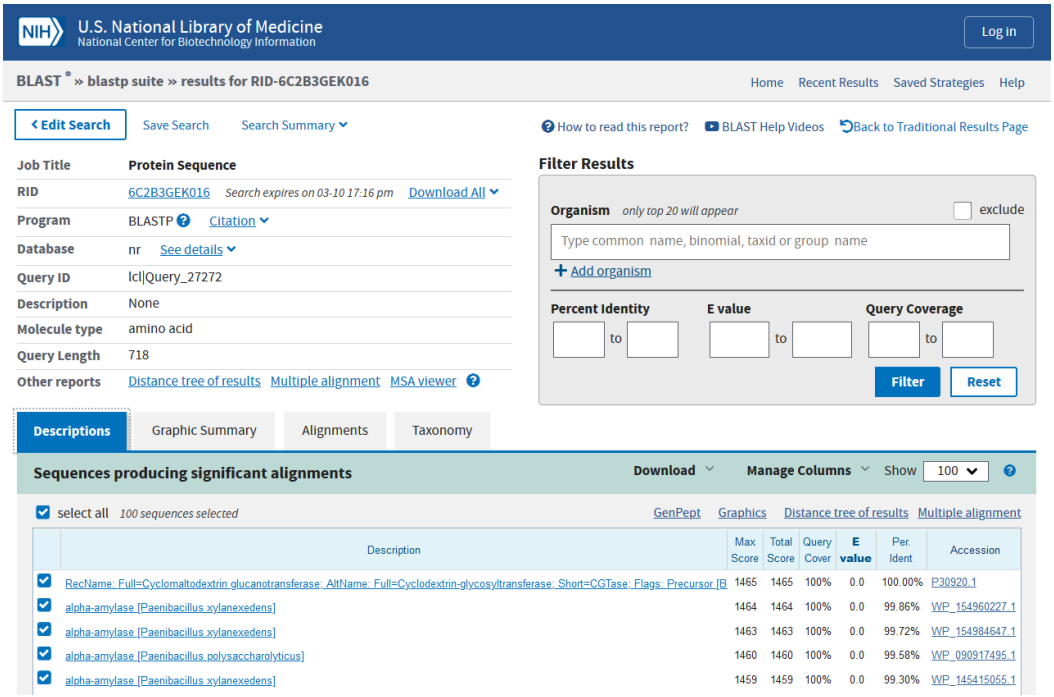


- výber parametra maximálneho počtu zobrazených sekvencií:

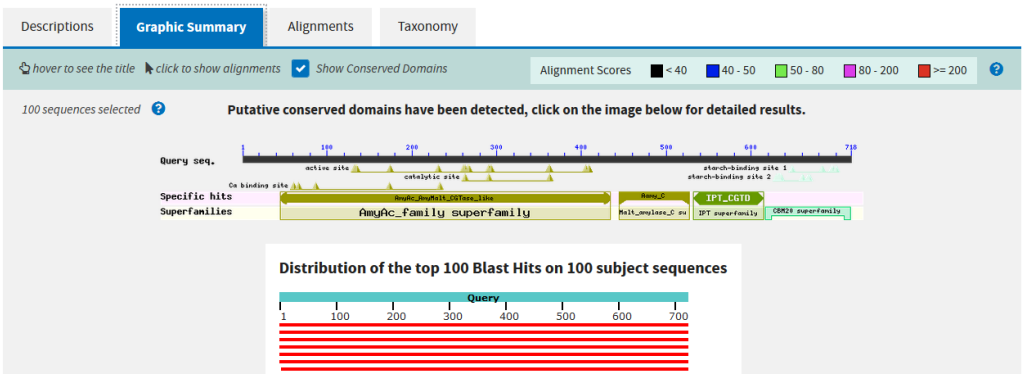


Výsledky: (i) opisy; (ii) grafický súhrn; (iii) zrovnania; a (iv) taxonómia:

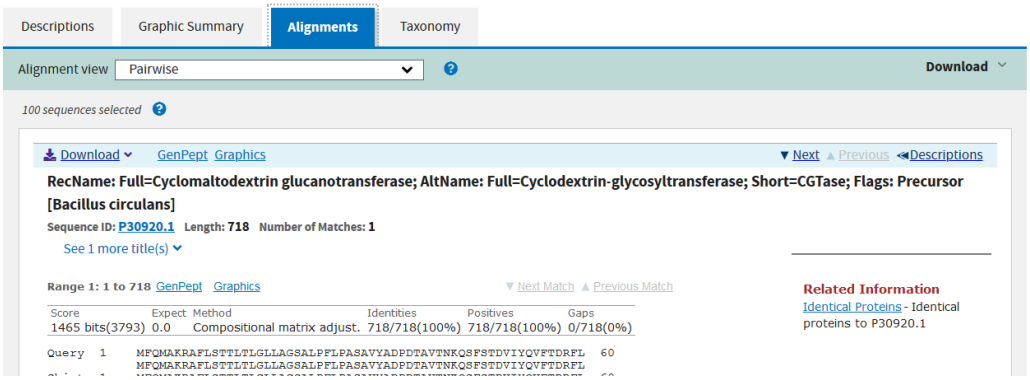
(i)



(ii)

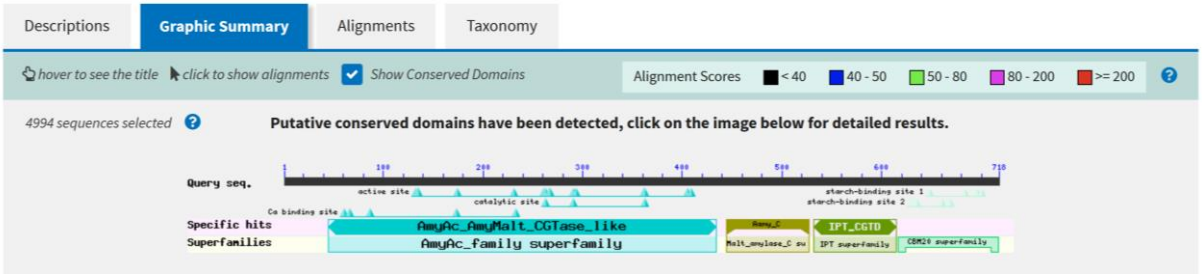


(iii)

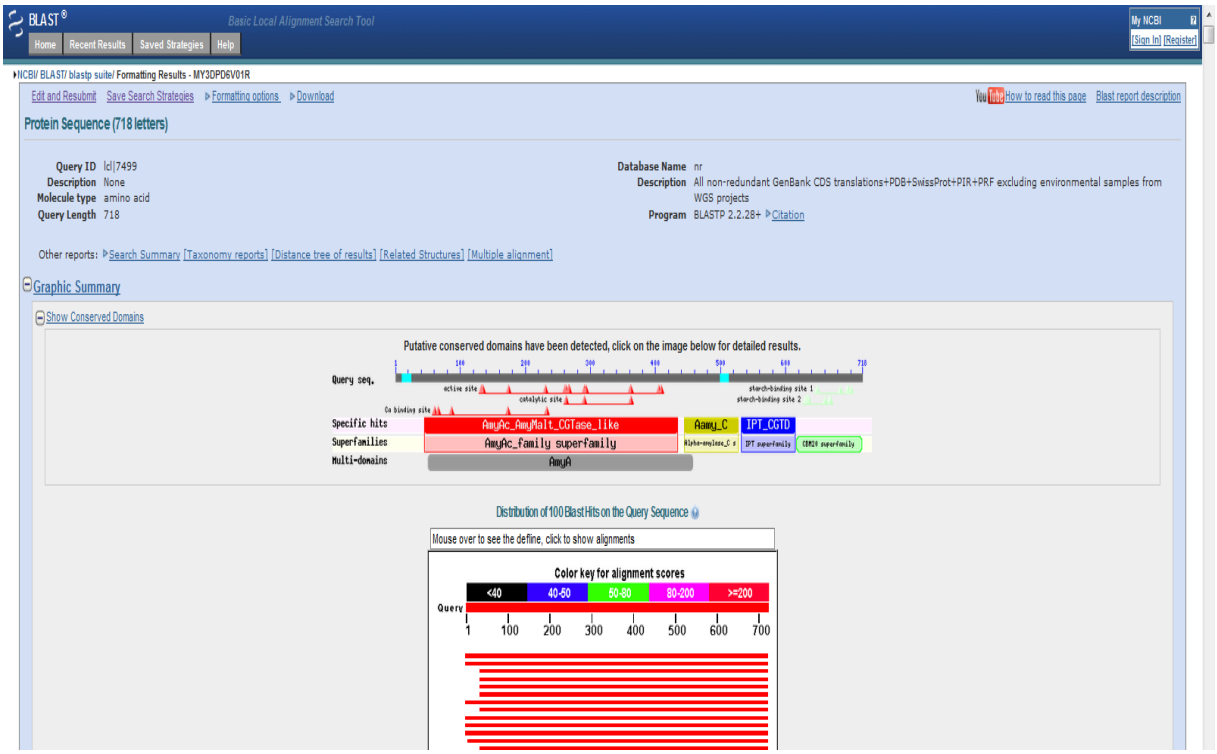


(iv)

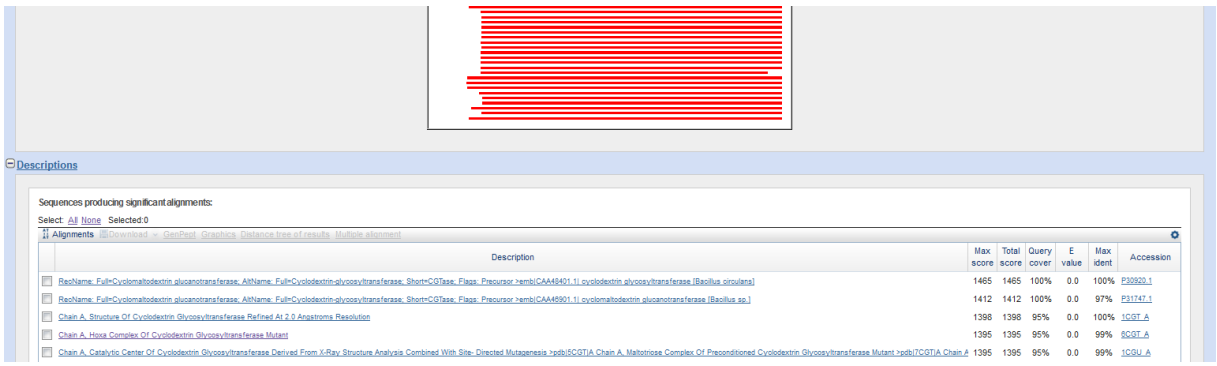
Identifikované konzervované domény v rámci grafického súhrnu:



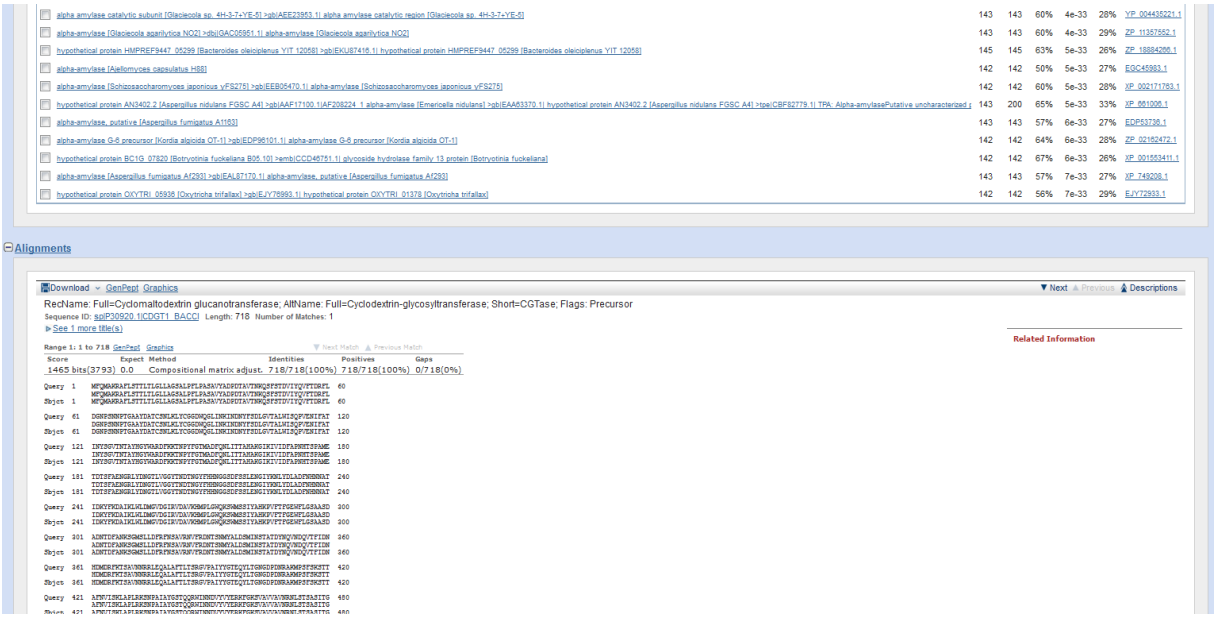
Grafický súhrn – identifikované konzervované domény a skóre zrovnania:



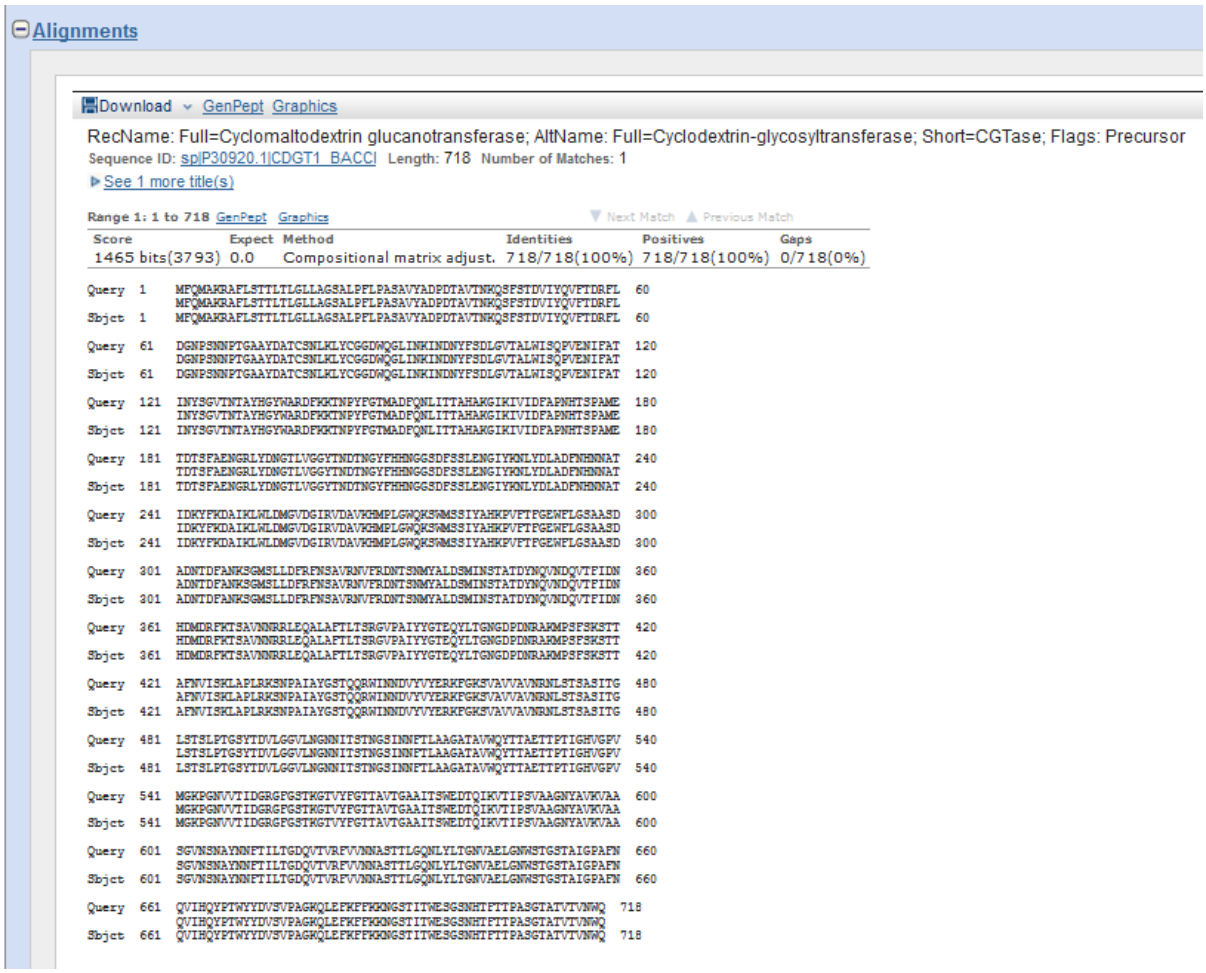
Koniec grafického súhrnu, začiatok opisov a v nich prvý zásah:



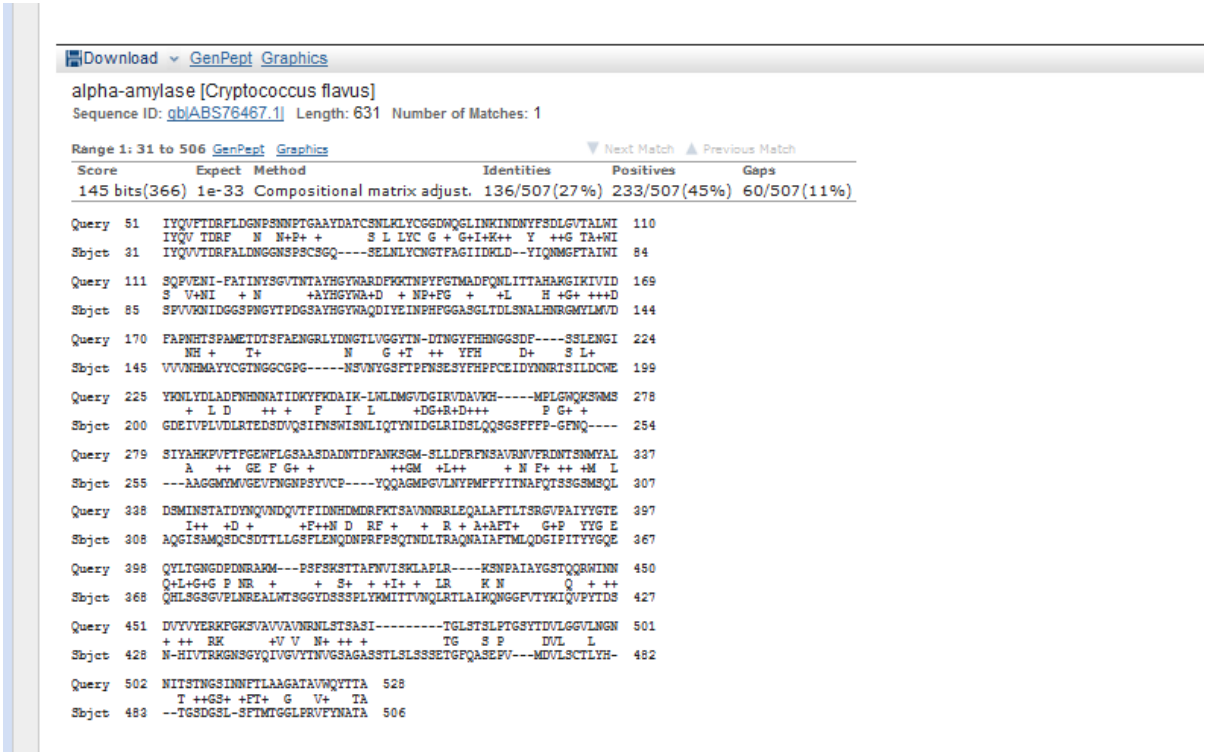
Koniec opisov, začiatok zrovnaní a v nich prvý zásah:



Prvý zásah v zrovnaniach – zrovnanie query samej so sebou:



Na porovnanie – n-tý zásah (vo všeobecnosti):



Download GenPept Graphics

alpha-amylase [Cryptococcus flavus]

Sequence ID: gb|ABS76467.1| Length: 631 Number of Matches: 1

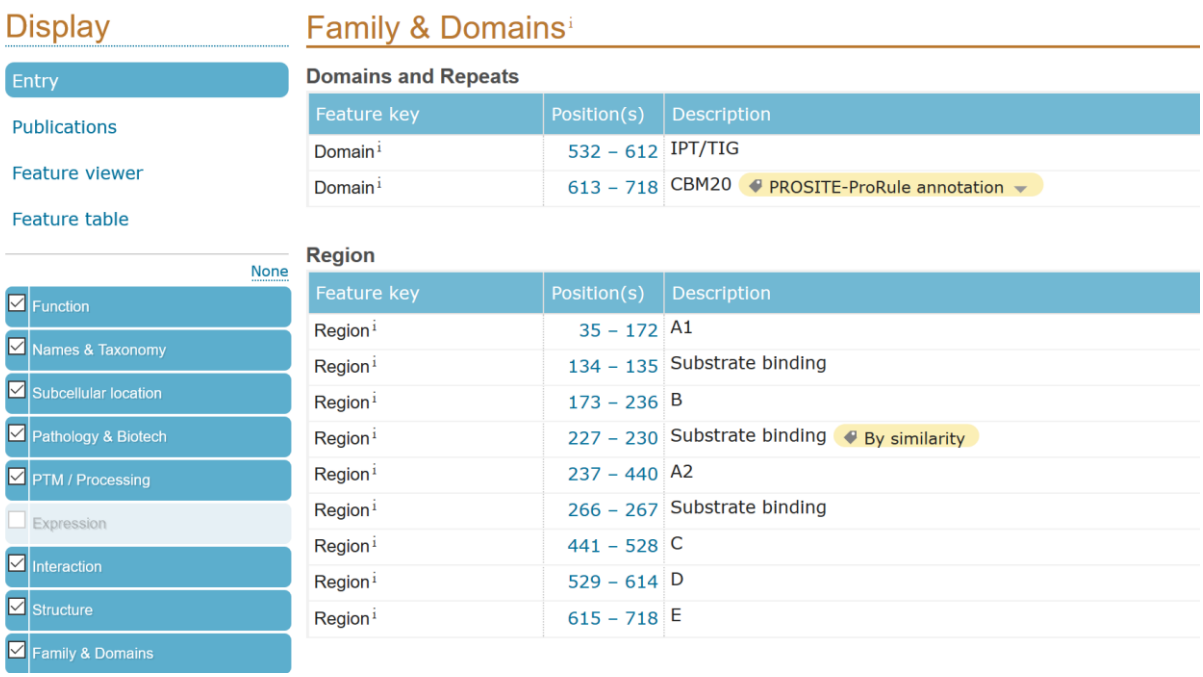
Range 1: 31 to 506 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
145 bits(366)	1e-33	Compositional matrix adjust.	136/507(27%)	233/507(45%)	60/507(11%)

```
Query  51   IYQVFTDRFLDGNPSNNPTGAAYDATCSNLKLYCGGDWQGLINKINDNYFSDLGVTALWI  110
            IYQV TDRF   N  N+P+ +      S L LYC G + G+I+K++  Y  ++G TA+WI
Sbjct  31   IYQVVTDRFALDNGGNSPSCSGQ----SELNLYCNGTFAGIIDKLD--YIQNMGFTAIWI  84

Query  111  SQPVENI-FATINYSGVTNTAYHGYWARDFKKTNPYFGTMADFQNLITTAHAKGIKIVID  169
            S  V+NI   + N      +AYHGYWA+D  + NP+PG  +   +L    H +G+ +++D
Sbjct  85   SFVVKNIDGGSPNGYTPDGSAYHGYWAQDIYEINPHFGGASGLTDLSNALHNRGMYLMVD  144

Query  170  FAPNHTSPAMETDTSFAENGRLYDNGTLVGGYTN-DTNGYFHHNGGSDF----SSLENGI  224
               NH +    T+           N    G +T  ++  YFH     D+    S L+
Sbjct  145  VVVNHMAYYCGTNGGCGPG-----NSVNYGSFTPFNSESYFHPFCEIDYNNRTSILDCWE  199

Query  225  YKNLYDLADFNHNNATIDKYFKDAIK-LWLDMGVDGIRVDAVKH-----MPLGWQKSWMS  278
              + L D    ++ +   F   I  L      +DG+R+D+++        P G+ +
Sbjct  200  GDEIVPLVDLRTEDSDVQSIFNSWISNLIQTYNIDGLRIDSLQQSGSFFFP-GFNQ----  254

Query  279  SIYAHKPVFTFGEWFLGSAASDADNTDFANKSGM-SLLDFRFNSAVRNVFRDNTSNMYAL  337
             A    ++  GE F G+ +          ++GM  +L++      + N F+ ++ +M  L
Sbjct  255  ---AAGGMYMVGEVFNGNPSYVCP----YQQAGMPGVLNYPMFFYITNAFQTSSGSMSQL  307

Query  338  DSMINSTATDYNQVNDQVTFIDNHDMDRFKTSAVNNRRLEQALAFTLTSRGVPAIYYGTE  397
              I++   +D +          +F++N D  RF +     +  R + A+AFT+    G+P  YYG E
Sbjct  308  AQGISAMQSDCSDTTLLGSFLENQDNPRFPSQTNDLTRAQNAIAFTMLQDGIPITYYGQE  367

Query  398  QYLTGNGDPDNRAKM---PSFSKSTTAFNVISKLAPLR----KSNPAIAYGSTQQRWINN  450
            Q+L+G+G P NR  +     +  S+  + +I+ +   LR    K N              Q  + ++
Sbjct  368  QHLSGSGVPLNREALWTSGGYDSSSPLYKMITTVNQLRTLAIKQNGGFVTYKIQVPYTDS  427

Query  451  DVYVYERKFGKSVAVVAVNRNLSTSASI---------TGLSTSLPTGSYTDVLGGVLNGN  501
            + ++  RK      +V V  N+ ++ +          TG   S P     DVL   L
Sbjct  428  N-HIVTRKGNSGYQIVGVYTNVGSAGASSTLSLSSSETGFQASEPV---MDVLSCTLYH-  482

Query  502  NITSTNGSINNFTLAAGATAVWQYTTA  528
              T ++GS+ +FT+  G   V+   TA
Sbjct  483  --TGSDGSL-SFTMTGGLPRVFYNATA  506
```

Blastovanie s jednotlivou doménou proteínu:

- záznam z databázy UniProt; doména E – región 615-718



Display

Entry

Publications

Feature viewer

Feature table

None

Function

Names & Taxonomy

Subcellular location

Pathology & Biotech

PTM / Processing

Expression

Interaction

Structure

Family & Domains

Family & Domains

Domains and Repeats

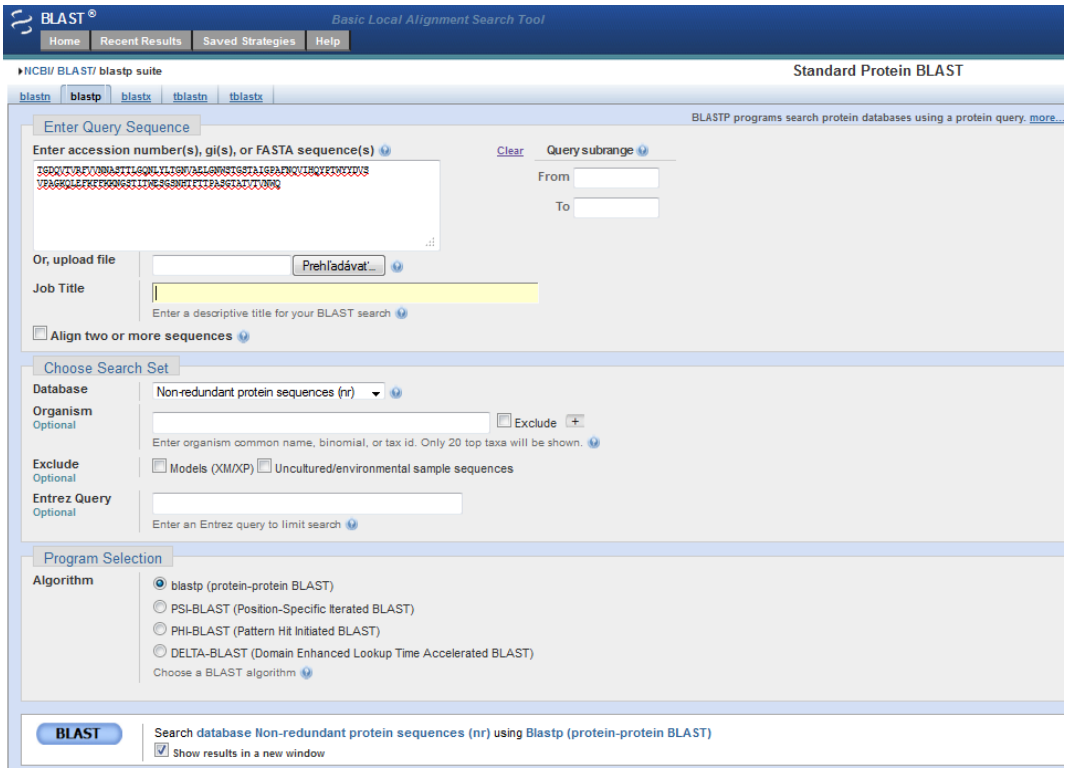
Feature key	Position(s)	Description
Domain	532 – 612	IPT/TIG
Domain	613 – 718	CBM20 PROSITE-ProRule annotation

Region

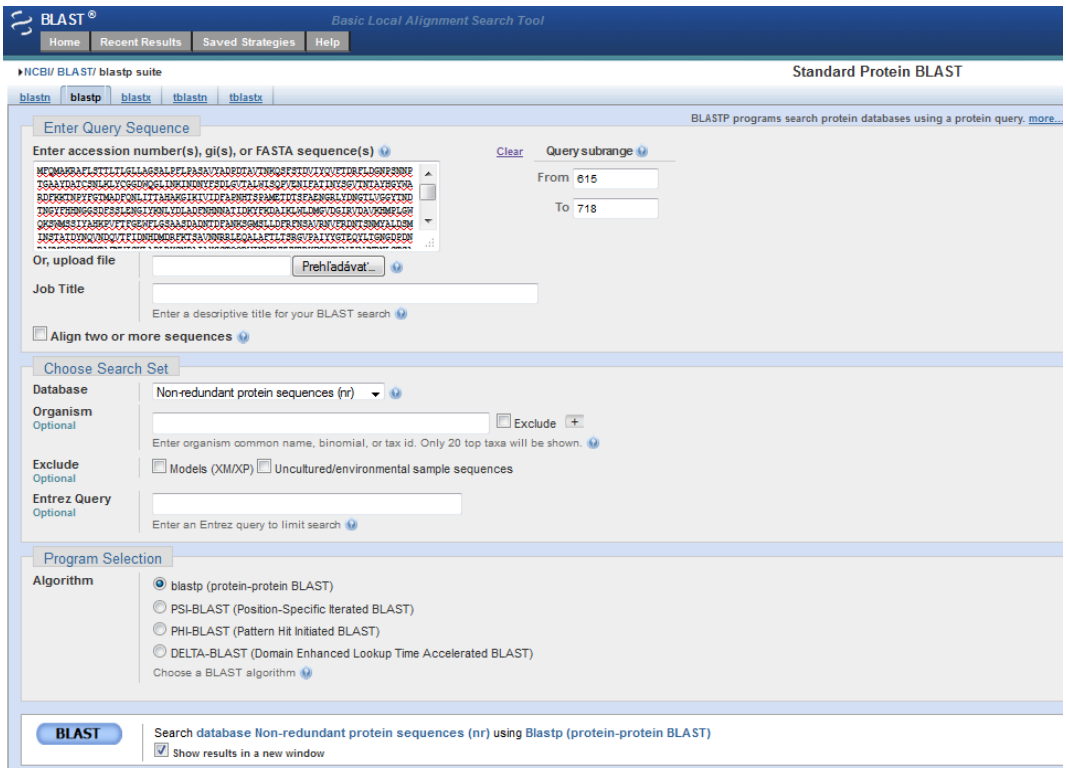
Feature key	Position(s)	Description
Region	35 – 172	A1
Region	134 – 135	Substrate binding
Region	173 – 236	B
Region	227 – 230	Substrate binding By similarity
Region	237 – 440	A2
Region	266 – 267	Substrate binding
Region	441 – 528	C
Region	529 – 614	D
Region	615 – 718	E

Vloženie query pri neúplnej sekvencii – dve možnosti postupu:

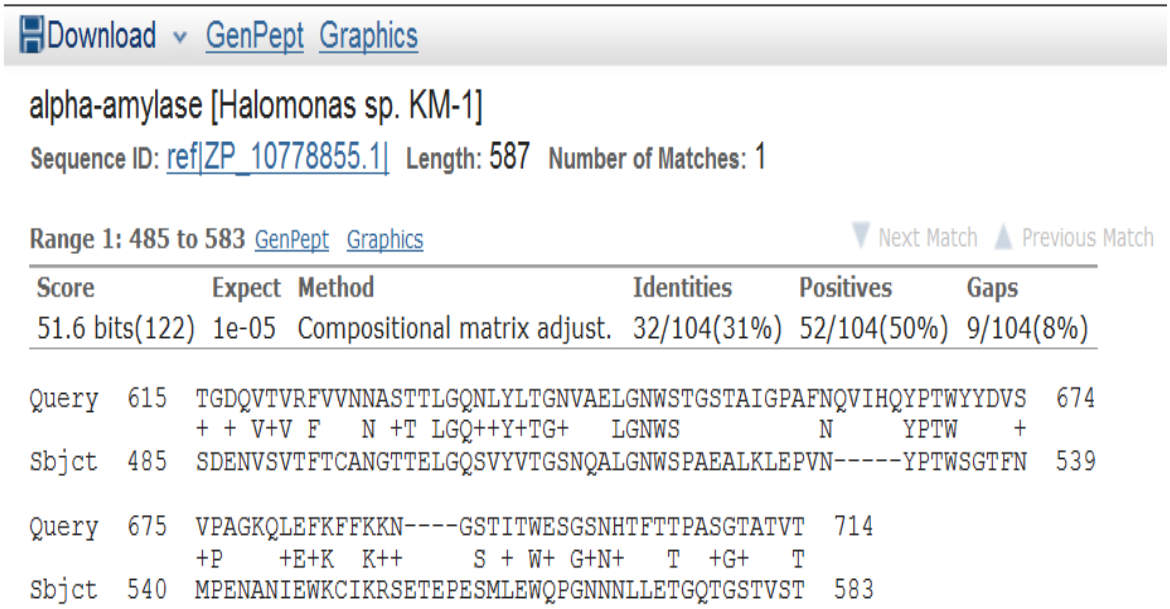
- vloží sa len konkrétny úsek sekvencie (napr. región 615-718)



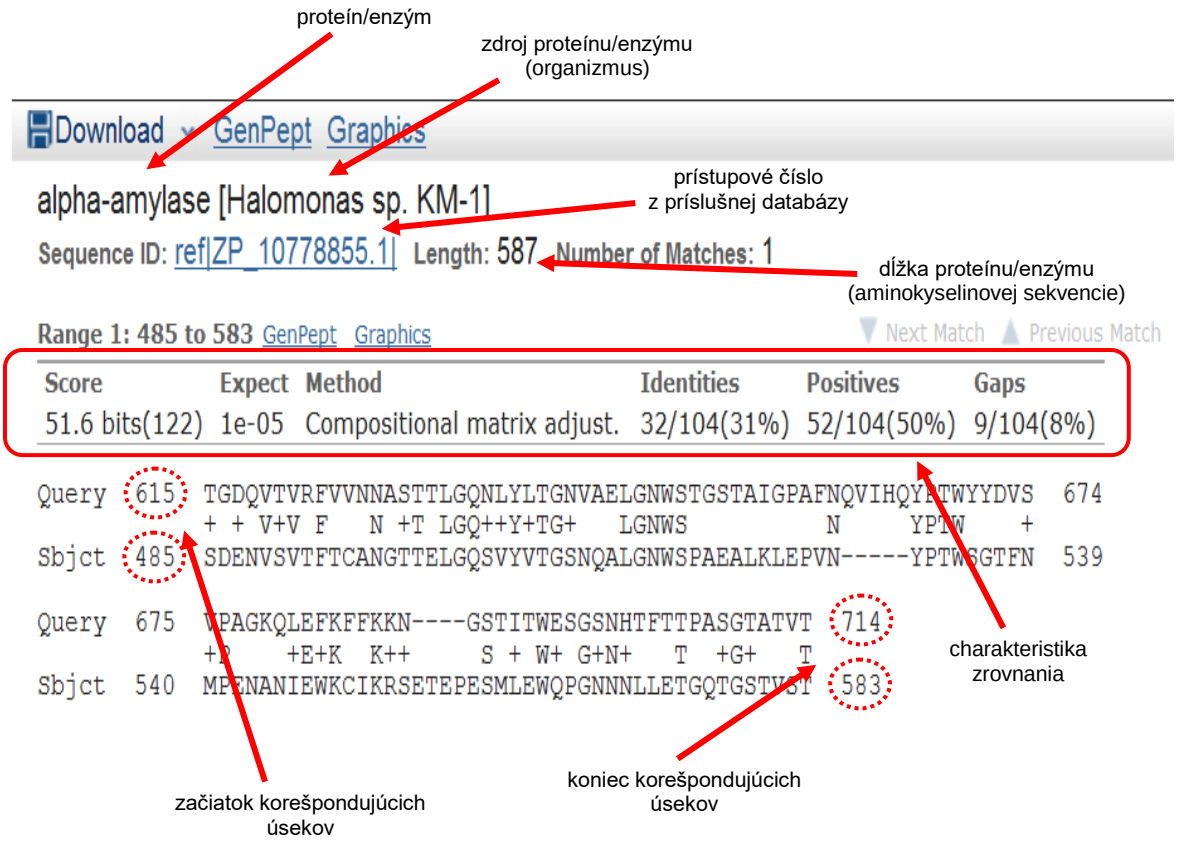
- alebo sa vloží celá sekvencie a vyznačia sa hranice („Query subrange“)



Detailný popis konkrétneho výsledku – n-tý zásah:

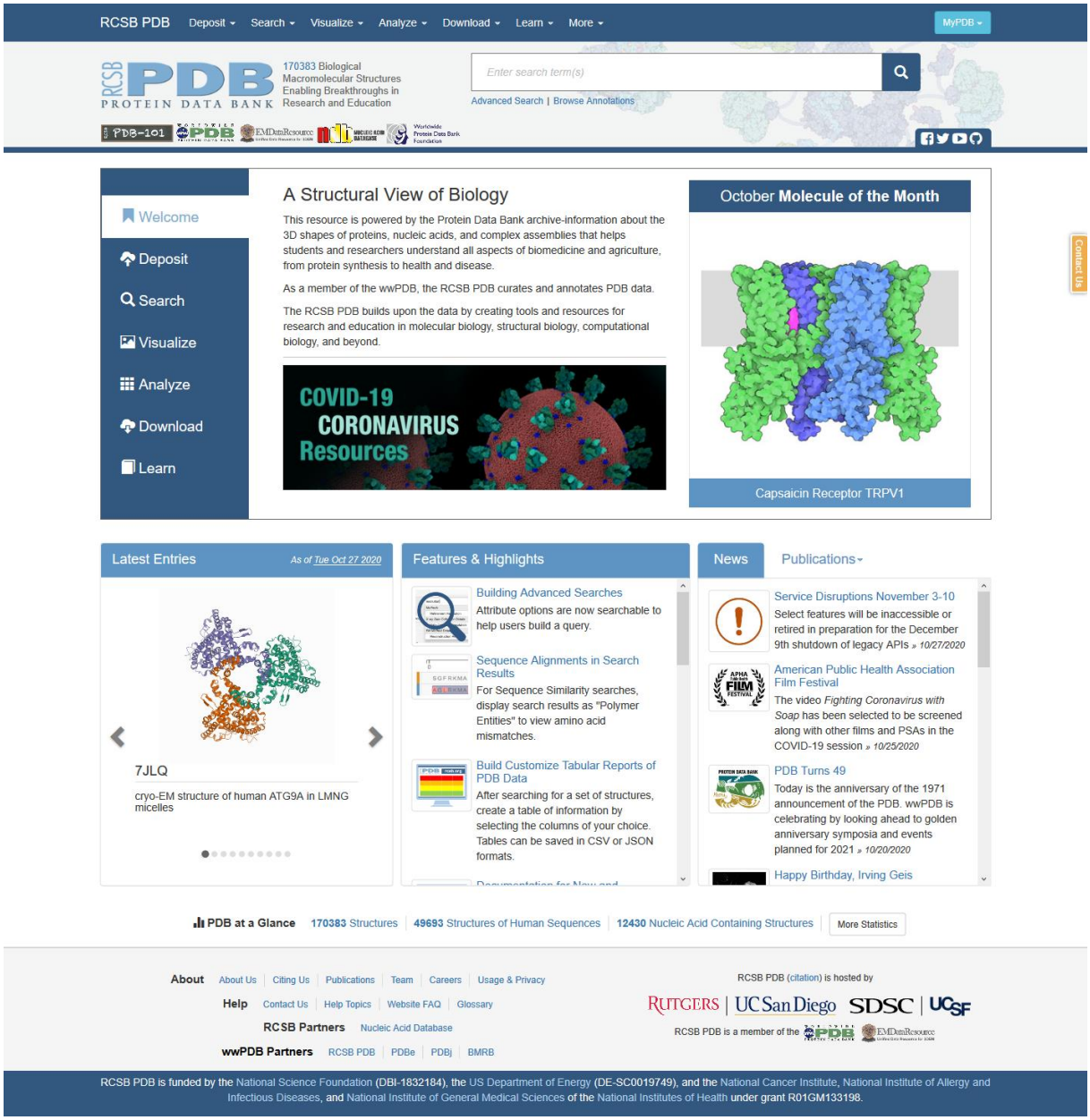


Možno povedať, že blastovaná doména (úsek sekvencie 615-718), resp. jej potenciálny homológ, sa nachádza v α-amyláze z *Halomonas* sp. KM-1 na C-konci proteínu, keďže korešpondujúce úseky sú 615-714 (takmer celý úsek) a 485-583 (dĺžka sekvencie „subject“ je 587 zvyškov).



6. PDB, modelovanie a porovnávanie štruktúr proteínov

Úvodná web-stránka databázy PDB – Protein Data Bank:
http://www.rcsb.org/



Protein Data Bank (PDB) je databáza, v ktorej sú uchovávané vyriešené terciárne štruktúry najmä proteínov, ale aj ďalších veľkých biologických molekúl (DNA a RNA). PDB vznikla v roku 1971 v rámci Brookhaven National Laboratory (USA) ako zdroj digitálnych dát s otvoreným prístupom v rámci celej biológie a medicíny. Do dnešnej internetovej podoby prešla

v roku 2003 ako tzv. „Worldwide PDB“, ktorá spravuje celý archív PDB a zabezpečuje, že celá databáza je voľne a verejne dostupná pre všetkých užívateľov bez obmedzeia. V súčasnosti predstavuje jeden z hlavných globálnych zdrojov experimentálnych dát, ktoré sú nevyhnutné pre vedecké bádanie a objavy.

Vyriešiť terciárnu štruktúru proteínu znamená vyriešiť – t.j. poznať – koordináty v priestore (x, y, z) – t.j. pozíciu – pre všetky atómy, ktoré tvoria molekulu proteínu. Najdôležitejšie sú 4 atómy – dusík (z -NH_2 skupiny), alfa-uhlík a karboxylový uhlík s kyslíkom (z -COOH skupiny), ktoré tvoria kostru polypeptidového reťazca. Samozrejme, pre presné určenie terciárnej štruktúry proteínu je potrebné určiť koordináty aj ostatných atómov z postranných reťazcov všetkých aminokyselinových zvyškov. Presnosť vyriešenia štruktúry proteínu – tzv. rozlíšenie – sa udáva v jednotkách „Å“ (10^{-10} m). Veľmi dobré (detailné) štruktúry proteínov sú vyriešené pri rozlíšení 1-2 Å; pri rozlíšení pod 1 Å sa stávajú „viditeľné“ aj atómy vodíka, ktoré sú príliš malé a ich pozícia sa pri horšom rozlíšení (t.j. vyššeej hodnote „Å“) nedá určiť. Treba si uvedomiť, že štruktúra proteínu je určená (vyriešená) tým lepšie (presnejšie), čím je hodnota rozlíšenia udaná v „Å“ nižšia; samotné rozlíšenie je tým väčšie, čím je hodnota „Å“ nižšia.

```
ATOM      1  N   ALA A   1      -8.065  17.989 -10.906  1.00  0.00           N
ATOM      2  CA  ALA A   1      -7.235  16.799 -11.163  1.00  0.00           C
ATOM      3  C   ALA A   1      -8.117  15.784 -10.473  1.00  0.00           C
ATOM      4  O   ALA A   1      -8.999  16.398  -9.862  1.00  0.00           O
ATOM      5  CB  ALA A   1      -6.004  16.975 -10.304  1.00  0.00           C
```

Obr. 6.1. Ukážka zápisu koordinát pre aminokyselinu Ala1 v reťazci „A“ α-amylázy z *Aspergillus oryzae*, ktoré sú v PDB uložené pod PDB kódom: 2TAA. Zápis pre alanín je tvorený 5 riadkami: N, Cα (CA), C, O a Cβ (CB).

Terciárne štruktúry sú v PDB uložené v súboroch, ktoré majú pridelené tzv. PDB kódy (podobne ako prístupové číslo v sekvenčných databázach; „Accession No.“). Minimálny počet riadkov v PDB súbore pre jednu aminokyselinu je 4, t.j. štyri atómy: N, Cα a C(O) uvedené vyššie (obr. 6.1), tvoriace kostru reťazca; to však platí pre glycín, ktorý nemá postranný reťazec. To znamená, že PDB súbory sú vo všeobecnosti veľmi veľké, keďže napr. priemerný proteín s 500 aminokyselinovými zvyškami musí obsahovať

viac ako 2 000 riadkov „ATOM“ s koordinátami; 2 000 riadkov by mal súbor vtedy, keby bol proteín tvorený iba samými glycínmi (t.j. iba 4 riadky na jeden zvyšok). Pritom v každom PDB súbore sa nachádza aj mnoho poznámok („REMARKS“); a väčšina štruktúr obsahuje na jednu molekulu proteínu niekoľko stoviek molekúl vody (sú vyriešené ako pozície atómu kyslíka z molekuly H_2O).

Detaily k terciárnej štruktúre proteínov, k organizácii PDB súboru proteínu v PDB, ako aj k teoretickým základom predikcií sekundárnej a terciárnej štruktúry proteínov sú uvedené v učebnom texte „*Proteínový dizajn*“, ktorý je dostupný online na adrese Univerzity sv. Cyrila a Metoda v Trnave: http://fpv.ucm.sk/images/ucebne_texty/Proteinovy_dizajn.pdf.

Obr. 6.2. Porovnanie nárastu dát v priebehu troch posledných rokov v sekvenčnej databáze GenBank a v štruktúrnej databáze PDB. Dochádza k neustále narastajúcemu rozdielu medzi poznaním sekvencií a poznaním štruktúr.

Možnosti predikcie terciárnych štruktúr proteínov poskytujú potenciál spomaliť nepriaznivý trend enormného nárastu sekvenčných dát oproti omnoho pomalšiemu nárastu štruktúrnych dát v PDB (obr. 6.2).

V nasledujúcej časti budú predstavené, resp. popísané jednotlivé praktické kroky pri získaní vhodného modelu terciárnej štruktúry proteínu, interpretácii výsledkov modelovania, ako aj následného porovnania štruktúry modelu s reálnou štruktúrou proteínu, ktorý je k nemu homologický, t.j. ktorý by mohol, ale aj nemusel byť použitý pri modelovaní ako jeho templát.

Pre predikciu štuktúry proteínov existujú rôzne servery; jedným z nich je server Phyre2, kde je možné získať model tzv. induktívnou metódou rozpoznania štruktúry („fold recognition“). Princípom je, že program vyhľadáva v PDB štruktúry homologických proteínov, ktoré majú čo najvyšší stupeň sekvenčnej podobnosti s proteínom, o predikciu štruktúry ktorého je záujem. Tieto homologické proteíny slúžia ako štruktúrne templáty (vzory). To znamená, že výsledkom – v prípade, že existuje dostatok experimentálne určených terciárnych štruktúr homologických proteínov – je teoreticky toľko modelov, koľko templátov je v PDB identifikovaných.

Phyre2 spravidla poskytuje maximálne 20 modelov; v prípade väčšieho počtu identifikovaných vhodných templátov program poskytne získané dáta bez štruktúrnych koordinát modelov.

Pre využitie a správnu interpretáciu výsledkov je dôležité vziať do úvahy niekoľko nasledujúcich poznámok:

(i) v stĺpci hneď vedľa poradového čísla („Template“) sú zapísané PDB kódy použitých templátov, napr. „d1cyga2“ a „d1kula“ znamená, že PDB kódy sú „1CYG“, resp. „1KUL“;

(ii) stĺpec „Alignment Coverage“ je veľmi dôležitý, pretože udáva, aká časť sekvencie bola modelovaná; čím bližšie k 100%, tým lepšie;

(iii) cez stĺpec „3D Model“ je možné získať koordináty (t.j. terciárnu štrukrúru) modelu sekvencie na danom prekryve s templátom (alignment coverage); je dôležité si štruktúru vhodne uložiť;

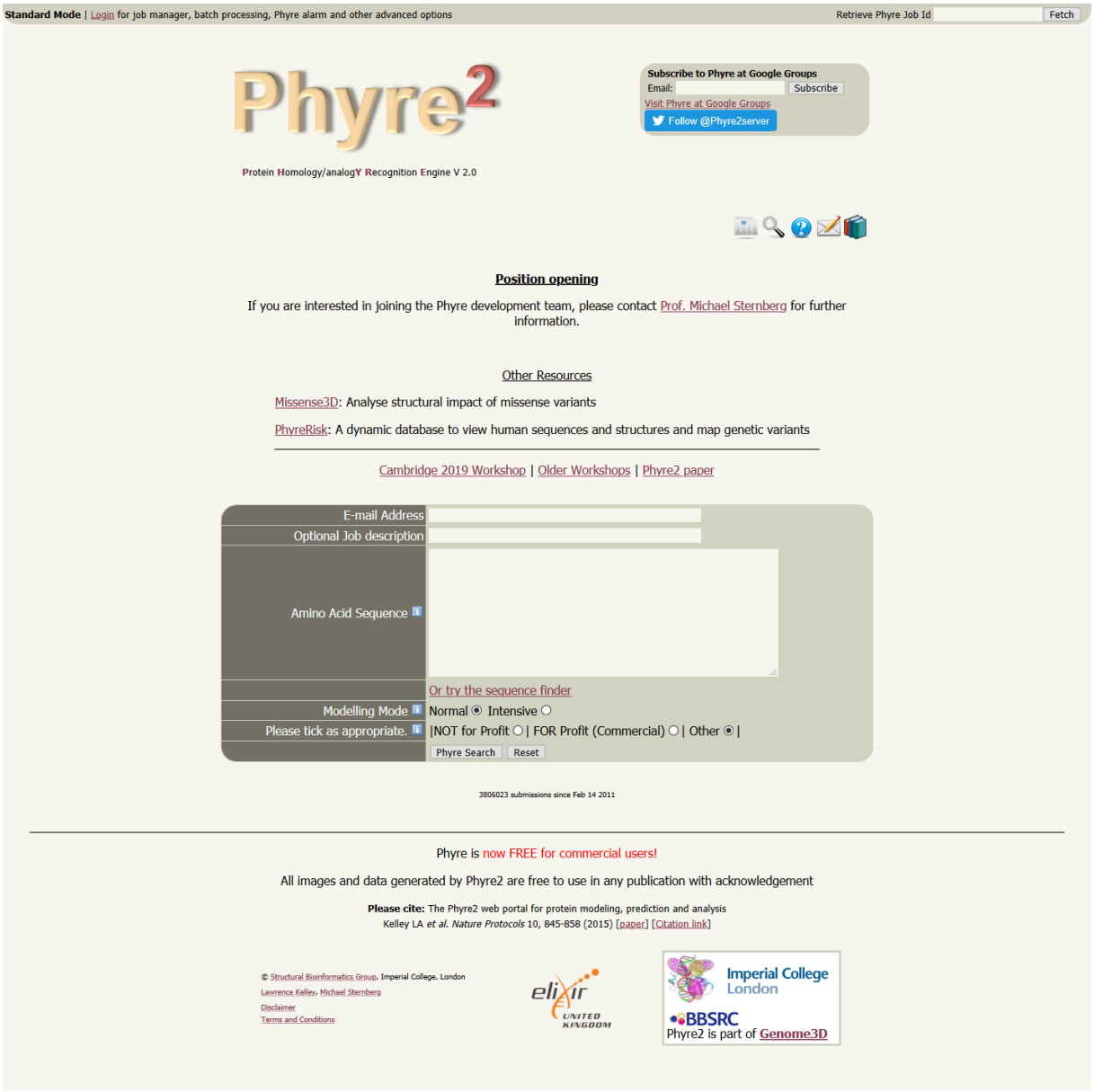
(iv) stĺpec „Confidence“ udáva pravdepodobnosť (0-100%), že templát a modelovaná sekvencia sú homológy; neudáva presnosť modelu;

(v) predposledný stĺpec „% i.d.“ znamená sekvenčnú identitu v rámci daného sekvenčného prekryvu; „% i.d.“ a „Alignment Coverage“ sú spolu najdôležitejšími hodnotami pre výber vhodného modelu;

(vi) posledný stĺpec obsahuje stručný popis templátu.

Úvodná web-stránka serveru Phyre2:

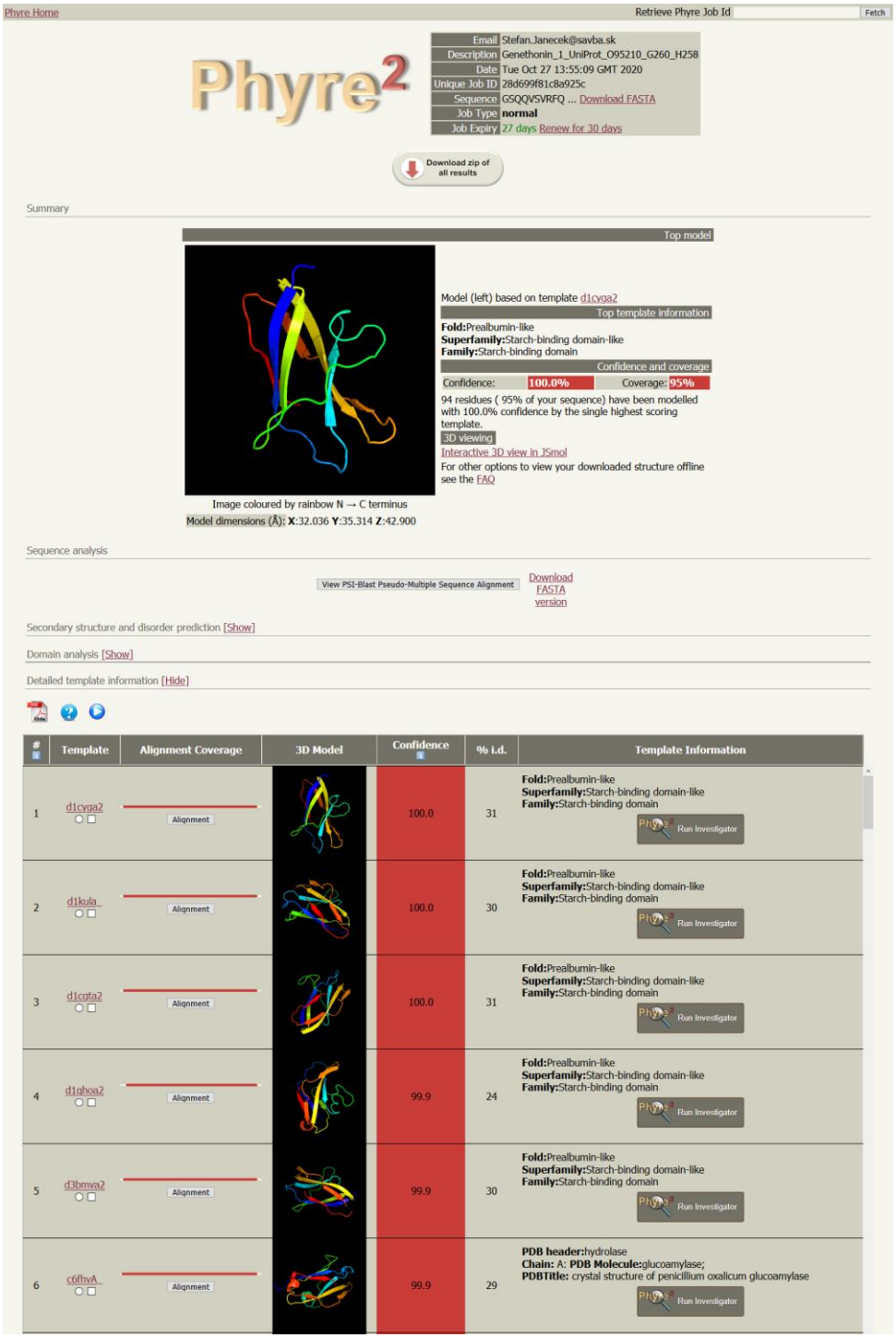
http://www.sbg.bio.ic.ac.uk/~phyre2/



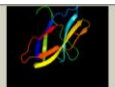
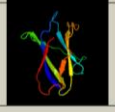
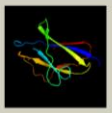
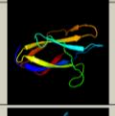
Na začiatku je potrebné vložiť aminokyselinovú sekvenciu proteínu (alebo jej časť), o predikciu terciárnej štruktúry ktorého je záujem. Je výhodné si túto predikciu nejako vhodne nazvať („Job description“) – napr. možno použiť názov proteínu, prístupové číslo z databázy GenBank alebo UniProt, prípadne rozsah aminokyselinovej sekvencie pri predikcii jej časti, apod. Pre bežné prípady úplne postačuje normálny spôsob modelovania („modelling mode“: „Normal“); t.j. nie „Intensive“.

Ako vidno nižšie, výsledky, ktoré poskytuje server Phyre2, sú v podstate veľmi rozsiahle. V tomto konkrétnom prípade bola modelovaná C-terminálna časť proteínu genetonín-1 (prístupové číslo z databázy UniProt: O95210) v úseku Gly260-His358. Táto časť korešponduje so škrob-viažucou doménou z rodiny CBM20 („carbohydrate-binding module family 20“) mikrobiálnych amylolytických enzýmov.

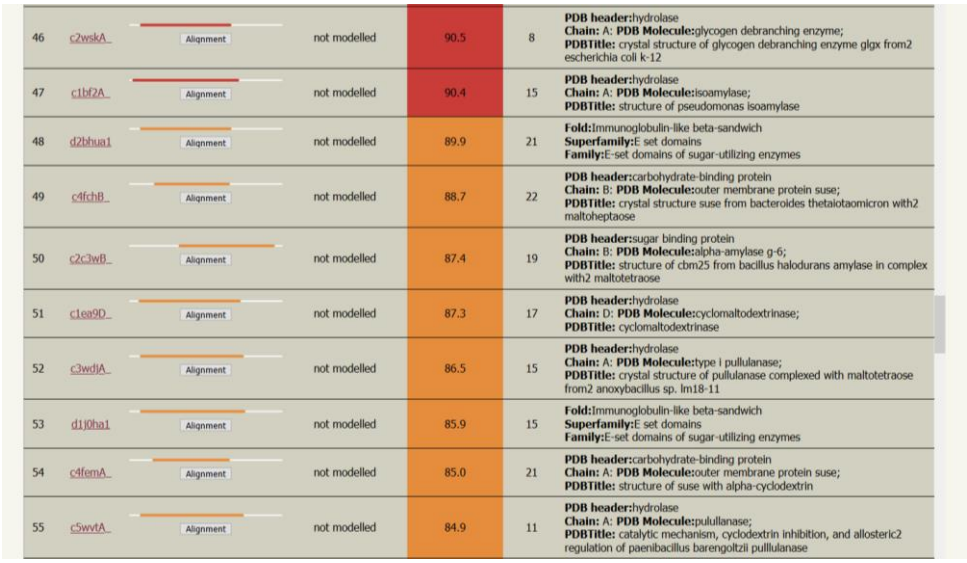
Výsledky modelovania zo serveru Phyre2:



Výsledky modelovania zo serveru Phyre2 (pokračovanie):

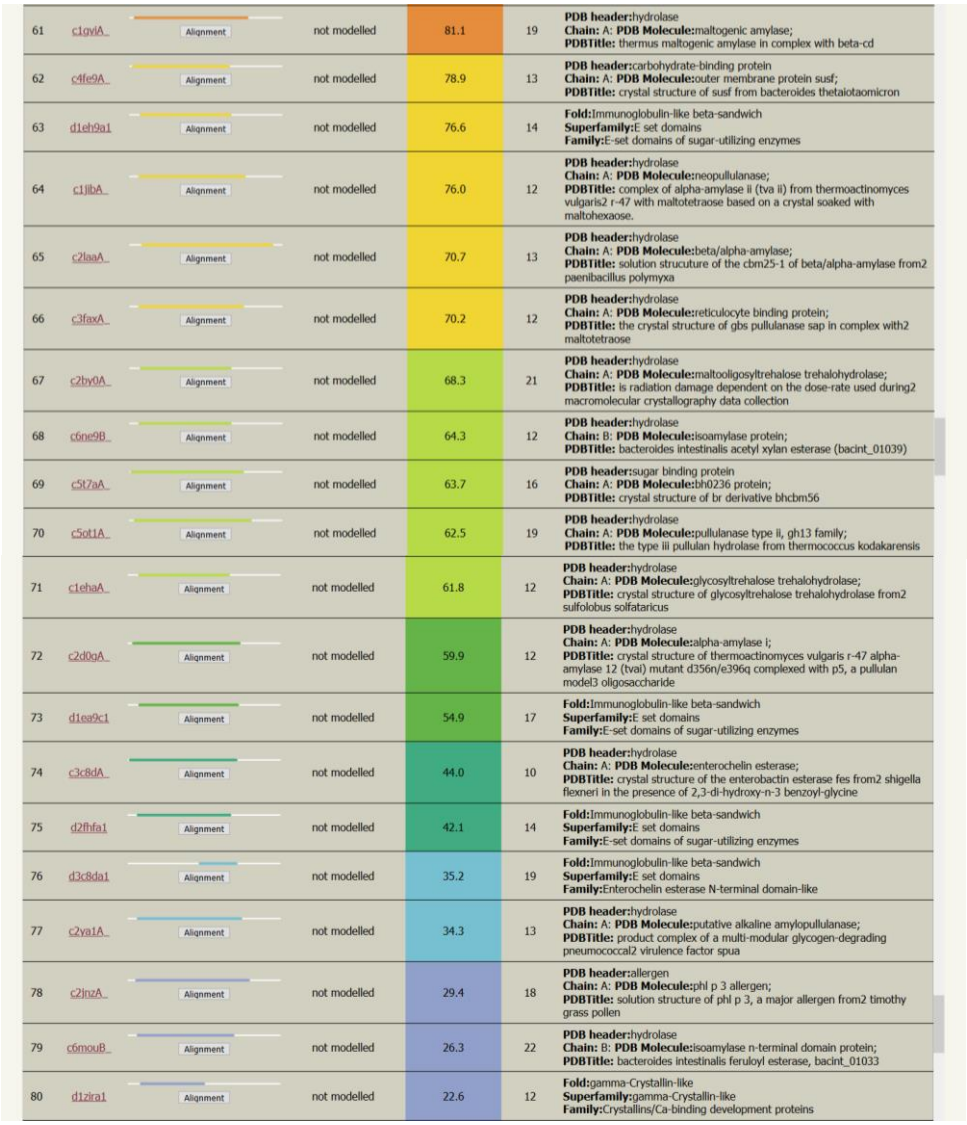
7	d1cda2	Alignment		99.9	32	**Superfamily:**Starch-binding domain-like **Family:**Starch-binding domain Run Investigator
8	d1pama2	Alignment		99.9	31	**Fold:**Prealbumin-like **Superfamily:**Starch-binding domain-like **Family:**Starch-binding domain Run Investigator
9	c2vn4A	Alignment		99.9	29	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**glucoamylase; **PDBTitle:** glycoside hydrolase family 15 glucoamylase from hypocrea jecorina Run Investigator
10	c6fhwB	Alignment		99.9	29	**PDB header:**hydrolase **Chain:** B: **PDB Molecule:**glucoamylase p; **PDBTitle:** structure of hormoconis resinae glucoamylase Run Investigator
11	d1vema1	Alignment		99.9	19	**Fold:**Prealbumin-like **Superfamily:**Starch-binding domain-like **Family:**Starch-binding domain Run Investigator
12	c2z0bB	Alignment		99.9	21	**PDB header:**hydrolase **Chain:** B: **PDB Molecule:**putative glycerophosphodiester phosphodiesterase 5; **PDBTitle:** crystal structure of cbm20 domain of human putative2 glycerophosphodiester phosphodiesterase 5 (kiaa1434) Run Investigator
13	c1f0yD	Alignment		99.7	22	**Chain:** D: **PDB Molecule:**beta-amylase; **PDBTitle:** beta-amylase from bacillus cereus var. mycoides in complex with2 glucose Run Investigator
14	c4rkkA	Alignment		99.2	26	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**laforin; **PDBTitle:** structure of a product bound phosphatase Run Investigator
15	c1cygA	Alignment		98.8	29	**PDB header:**glycosyltransferase **Chain:** A: **PDB Molecule:**cyclodextrin glucanotransferase; **PDBTitle:** cyclodextrin glucanotransferase (e.c.2.4.1.19) (cgtase) Run Investigator
16	c1qhoA	Alignment		98.7	25	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**alpha-amylase; **PDBTitle:** five-domain alpha-amylase from bacillus stearothermophilus,2 maltose/acarbose complex Run Investigator
17	c3nmeA	Alignment		98.5	24	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**sex4 glucan phosphatase; **PDBTitle:** structure of a plant phosphatase Run Investigator
18	c3bmwA	Alignment		98.3	28	**PDB header:**transferase **Chain:** A: **PDB Molecule:**cyclomaltodextrin glucanotransferase; **PDBTitle:** cyclodextrin glycosyl transferase from thermoanerobacterium2 thermosulfurigenes em1 mutant s77p complexed with a maltoheptaose3 inhibitor Run Investigator
19	c4jcmA	Alignment		98.3	28	**PDB header:**transferase **Chain:** A: **PDB Molecule:**cyclodextrin glucanotransferase; **PDBTitle:** crystal structure of gamma-cgtase from alkalophilic bacillus clarkii2 at 1.65 angstrom resolution Run Investigator
20	d2f15a1	Alignment		98.1	21	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**AMPK-beta glycogen binding domain-like Run Investigator
21	c4jclA	Alignment	not modelled	98.1	27	**PDB header:**transferase **Chain:** A: **PDB Molecule:**cyclomaltodextrin glucanotransferase; **PDBTitle:** crystal structure of alpha-cgt from paenibacillus macerans at 1.72 angstrom resolution
22	c1tcmB	Alignment	not modelled	98.0	26	**PDB header:**glycosyltransferase **Chain:** B: **PDB Molecule:**cyclodextrin glycosyltransferase; **PDBTitle:** cyclodextrin glycosyltransferase w616a mutant from bacillus2 circulans strain 251
23	c2qlvB	Alignment	not modelled	97.7	23	**PDB header:**transferase/protein binding **Chain:** B: **PDB Molecule:**protein sip2; **PDBTitle:** crystal structure of the heterotrimer core of the s. cerevisiae ampk2 homolog snf1
24	d2qlvb1	Alignment	not modelled	97.7	18	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**AMPK-beta glycogen binding domain-like
25	d1z0na1	Alignment	not modelled	97.6	13	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**AMPK-beta glycogen binding domain-like
26	d1z0mb1	Alignment	not modelled	97.5	18	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**AMPK-beta glycogen binding domain-like
27	c4bzvC	Alignment	not modelled	97.1	16	**PDB header:**transferase **Chain:** C: **PDB Molecule:**1,4-alpha-glucan-branching enzyme; **PDBTitle:** crystal structure of human glycogen branching enzyme (gbe1)

Výsledky modelovania zo serveru Phyre2 (pokračovanie, prerušenie: 27→46):



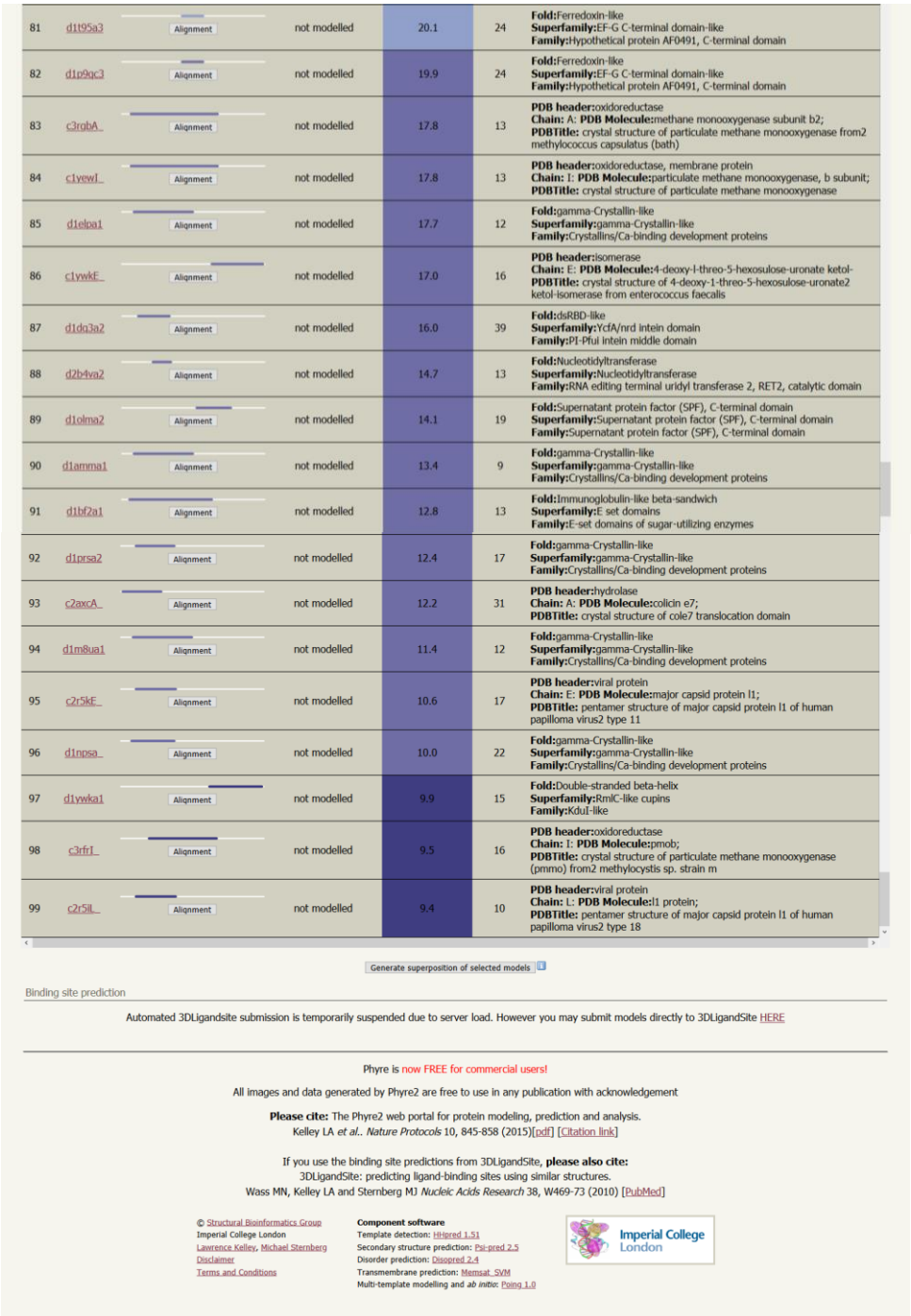
46	c2wskA_	Alignment	not modelled	90.5	8	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**glycogen debranching enzyme; **PDBTitle:** crystal structure of glycogen debranching enzyme glgx from2 escherichia coli k-12
47	c1bf2A_	Alignment	not modelled	90.4	15	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**isoamylase; **PDBTitle:** structure of pseudomonas isoamylase
48	d2bhua1	Alignment	not modelled	89.9	21	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**E-set domains of sugar-utilizing enzymes
49	c4fchB_	Alignment	not modelled	88.7	22	**PDB header:**carbohydrate-binding protein **Chain:** B: **PDB Molecule:**outer membrane protein suse; **PDBTitle:** crystal structure suse from bacteroides thetaiotaomicron with2 maltoheptaose
50	c2c3wB_	Alignment	not modelled	87.4	19	**PDB header:**sugar binding protein **Chain:** B: **PDB Molecule:**alpha-amylase g-6; **PDBTitle:** structure of cbm25 from bacillus halodurans amylase in complex with2 maltotetraose
51	c1ea9D_	Alignment	not modelled	87.3	17	**PDB header:**hydrolase **Chain:** D: **PDB Molecule:**cyclomaltodextrinase; **PDBTitle:** cyclomaltodextrinase
52	c3wdjA_	Alignment	not modelled	86.5	15	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**type i pullulanase; **PDBTitle:** crystal structure of pullulanase complexed with maltotetraose from2 anoxybacillus sp. lm18-11
53	d1j0ha1	Alignment	not modelled	85.9	15	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**E-set domains of sugar-utilizing enzymes
54	c4femA_	Alignment	not modelled	85.0	21	**PDB header:**carbohydrate-binding protein **Chain:** A: **PDB Molecule:**outer membrane protein suse; **PDBTitle:** structure of suse with alpha-cyclodextrin
55	c5wvtA_	Alignment	not modelled	84.9	11	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**pululllanase; **PDBTitle:** catalytic mechanism, cyclodextrin inhibition, and allosteric2 regulation of paenibacillus barengoltzii pulllulanase

Výsledky modelovania zo serveru Phyre2 (pokračovanie, prerušenie: 55→61):



61	c1gviA_	Alignment	not modelled	81.1	19	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**maltogenic amylase; **PDBTitle:** thermus maltogenic amylase in complex with beta-cd
62	c4fe9A_	Alignment	not modelled	78.9	13	**PDB header:**carbohydrate-binding protein **Chain:** A: **PDB Molecule:**outer membrane protein susf; **PDBTitle:** crystal structure of susf from bacteroides thetaiotaomicron
63	d1eh9a1	Alignment	not modelled	76.6	14	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**E-set domains of sugar-utilizing enzymes
64	c1jlbA_	Alignment	not modelled	76.0	12	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**neopullulanase; **PDBTitle:** complex of alpha-amylase ii (tva ii) from thermoactinomyces vulgaris2 r-47 with maltotetraose based on a crystal soaked with maltohexaose.
65	c2laaA_	Alignment	not modelled	70.7	13	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**beta/alpha-amylase; **PDBTitle:** solution strucuture of the cbm25-1 of beta/alpha-amylase from2 paenibacillus polymyxa
66	c3faxA_	Alignment	not modelled	70.2	12	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**reticulocyte binding protein; **PDBTitle:** the crystal structure of gbs pullulanase sap in complex with2 maltotetraose
67	c2by0A_	Alignment	not modelled	68.3	21	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**maltooligosyltrehalose trehalohydrolase; **PDBTitle:** is radiation damage dependent on the dose-rate used during2 macromolecular crystallography data collection
68	c6ne9B_	Alignment	not modelled	64.3	12	**PDB header:**hydrolase **Chain:** B: **PDB Molecule:**isoamylase protein; **PDBTitle:** bacteroides intestinalis acetyl xylan esterase (bacint_01039)
69	c5t7aA_	Alignment	not modelled	63.7	16	**PDB header:**sugar binding protein **Chain:** A: **PDB Molecule:**bh0236 protein; **PDBTitle:** crystal structure of br derivative bhcbm56
70	c5ot1A_	Alignment	not modelled	62.5	19	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**pullulanase type ii, gh13 family; **PDBTitle:** the type iii pullulan hydrolase from thermococcus kodakarensis
71	c1ehaA_	Alignment	not modelled	61.8	12	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**glycosyltrehalose trehalohydrolase; **PDBTitle:** crystal structure of glycosyltrehalose trehalohydrolase from2 sulfolobus solfataricus
72	c2d0gA_	Alignment	not modelled	59.9	12	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**alpha-amylase i; **PDBTitle:** crystal structure of thermoactinomyces vulgaris r-47 alpha-amylase 12 (tvai) mutant d356n/e396q complexed with p5, a pullulan model3 oligosaccharide
73	d1ea9c1	Alignment	not modelled	54.9	17	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**E-set domains of sugar-utilizing enzymes
74	c3c8dA_	Alignment	not modelled	44.0	10	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**enterochelin esterase; **PDBTitle:** crystal structure of the enterobactin esterase fes from2 shigella flexneri in the presence of 2,3-di-hydroxy-n-3 benzoyl-glycine
75	d2fhfa1	Alignment	not modelled	42.1	14	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**E-set domains of sugar-utilizing enzymes
76	d3c8da1	Alignment	not modelled	35.2	19	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**Enterochelin esterase N-terminal domain-like
77	c2ya1A_	Alignment	not modelled	34.3	13	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**putative alkaline amylopullulanase; **PDBTitle:** product complex of a multi-modular glycogen-degrading pneumococcal2 virulence factor spua
78	c2lnzA_	Alignment	not modelled	29.4	18	**PDB header:**allergen **Chain:** A: **PDB Molecule:**phl p 3 allergen; **PDBTitle:** solution structure of phl p 3, a major allergen from2 timothy grass pollen
79	c6mouB_	Alignment	not modelled	26.3	22	**PDB header:**hydrolase **Chain:** B: **PDB Molecule:**isoamylase n-terminal domain protein; **PDBTitle:** bacteroides intestinalis feruloyl esterase, bacint_01033
80	d1zira1	Alignment	not modelled	22.6	12	**Fold:**gamma-Crystallin-like **Superfamily:**gamma-Crystallin-like **Family:**Crystallins/Ca-binding development proteins

Výsledky modelovania zo serveru Phyre2 (pokračovanie; dokončenie):



81	d1t95a3	Alignment	not modelled	20.1	24	**Fold:**Ferredoxin-like **Superfamily:**EF-G C-terminal domain-like **Family:**Hypothetical protein AF0491, C-terminal domain
82	d1p9qc3	Alignment	not modelled	19.9	24	**Fold:**Ferredoxin-like **Superfamily:**EF-G C-terminal domain-like **Family:**Hypothetical protein AF0491, C-terminal domain
83	c3rqbA_	Alignment	not modelled	17.8	13	**PDB header:**oxidoreductase **Chain:** A: **PDB Molecule:**methane monooxygenase subunit b2; **PDBTitle:** crystal structure of particulate methane monooxygenase from2 methylococcus capsulatus (bath)
84	c1yewI_	Alignment	not modelled	17.8	13	**PDB header:**oxidoreductase, membrane protein **Chain:** I: **PDB Molecule:**particulate methane monooxygenase, b subunit; **PDBTitle:** crystal structure of particulate methane monooxygenase
85	d1elpa1	Alignment	not modelled	17.7	12	**Fold:**gamma-Crystallin-like **Superfamily:**gamma-Crystallin-like **Family:**Crystallins/Ca-binding development proteins
86	c1ywkE_	Alignment	not modelled	17.0	16	**PDB header:**isomerase **Chain:** E: **PDB Molecule:**4-deoxy-l-threo-5-hexosulose-uronate ketol- **PDBTitle:** crystal structure of 4-deoxy-1-threo-5-hexosulose-uronate2 ketol-isomerase from enterococcus faecalis
87	d1dq3a2	Alignment	not modelled	16.0	39	**Fold:**dsRBD-like **Superfamily:**YcfA/nrd intein domain **Family:**PI-Pfui intein middle domain
88	d2b4va2	Alignment	not modelled	14.7	13	**Fold:**Nucleotidyltransferase **Superfamily:**Nucleotidyltransferase **Family:**RNA editing terminal uridyl transferase 2, RET2, catalytic domain
89	d1olma2	Alignment	not modelled	14.1	19	**Fold:**Supernatant protein factor (SPF), C-terminal domain **Superfamily:**Supernatant protein factor (SPF), C-terminal domain **Family:**Supernatant protein factor (SPF), C-terminal domain
90	d1amma1	Alignment	not modelled	13.4	9	**Fold:**gamma-Crystallin-like **Superfamily:**gamma-Crystallin-like **Family:**Crystallins/Ca-binding development proteins
91	d1bf2a1	Alignment	not modelled	12.8	13	**Fold:**Immunoglobulin-like beta-sandwich **Superfamily:**E set domains **Family:**E-set domains of sugar-utilizing enzymes
92	d1prsa2	Alignment	not modelled	12.4	17	**Fold:**gamma-Crystallin-like **Superfamily:**gamma-Crystallin-like **Family:**Crystallins/Ca-binding development proteins
93	c2axcA_	Alignment	not modelled	12.2	31	**PDB header:**hydrolase **Chain:** A: **PDB Molecule:**colicin e7; **PDBTitle:** crystal structure of cole7 translocation domain
94	d1m8ua1	Alignment	not modelled	11.4	12	**Fold:**gamma-Crystallin-like **Superfamily:**gamma-Crystallin-like **Family:**Crystallins/Ca-binding development proteins
95	c2r5kE_	Alignment	not modelled	10.6	17	**PDB header:**viral protein **Chain:** E: **PDB Molecule:**major capsid protein l1; **PDBTitle:** pentamer structure of major capsid protein l1 of human papilloma virus2 type 11
96	d1npsa_	Alignment	not modelled	10.0	22	**Fold:**gamma-Crystallin-like **Superfamily:**gamma-Crystallin-like **Family:**Crystallins/Ca-binding development proteins
97	d1ywka1	Alignment	not modelled	9.9	15	**Fold:**Double-stranded beta-helix **Superfamily:**RmlC-like cupins **Family:**KduI-like
98	c3rfr1_	Alignment	not modelled	9.5	16	**PDB header:**oxidoreductase **Chain:** I: **PDB Molecule:**pmob; **PDBTitle:** crystal structure of particulate methane monooxygenase (pmmo) from2 methylocystis sp. strain m
99	c2r5iL_	Alignment	not modelled	9.4	10	**PDB header:**viral protein **Chain:** L: **PDB Molecule:**l1 protein; **PDBTitle:** pentamer structure of major capsid protein l1 of human papilloma virus2 type 18

Generate superposition of selected models

Binding site prediction

Automated 3DLigandsite submission is temporarily suspended due to server load. However you may submit models directly to 3DLigandSite HERE

Phyre is now FREE for commercial users!

All images and data generated by Phyre2 are free to use in any publication with acknowledgement

Please cite: The Phyre2 web portal for protein modeling, prediction and analysis.
Kelley LA *et al.*. *Nature Protocols* 10, 845-858 (2015)[pdf] [Citation link]

If you use the binding site predictions from 3DLigandSite, **please also cite:**
3DLigandSite: predicting ligand-binding sites using similar structures.
Wass MN, Kelley LA and Sternberg MJ *Nucleic Acids Research* 38, W469-73 (2010) [PubMed]

© Structural Bioinformatics Group
Imperial College London
Lawrence Kelley, Michael Sternberg
Disclaimer
Terms and Conditions

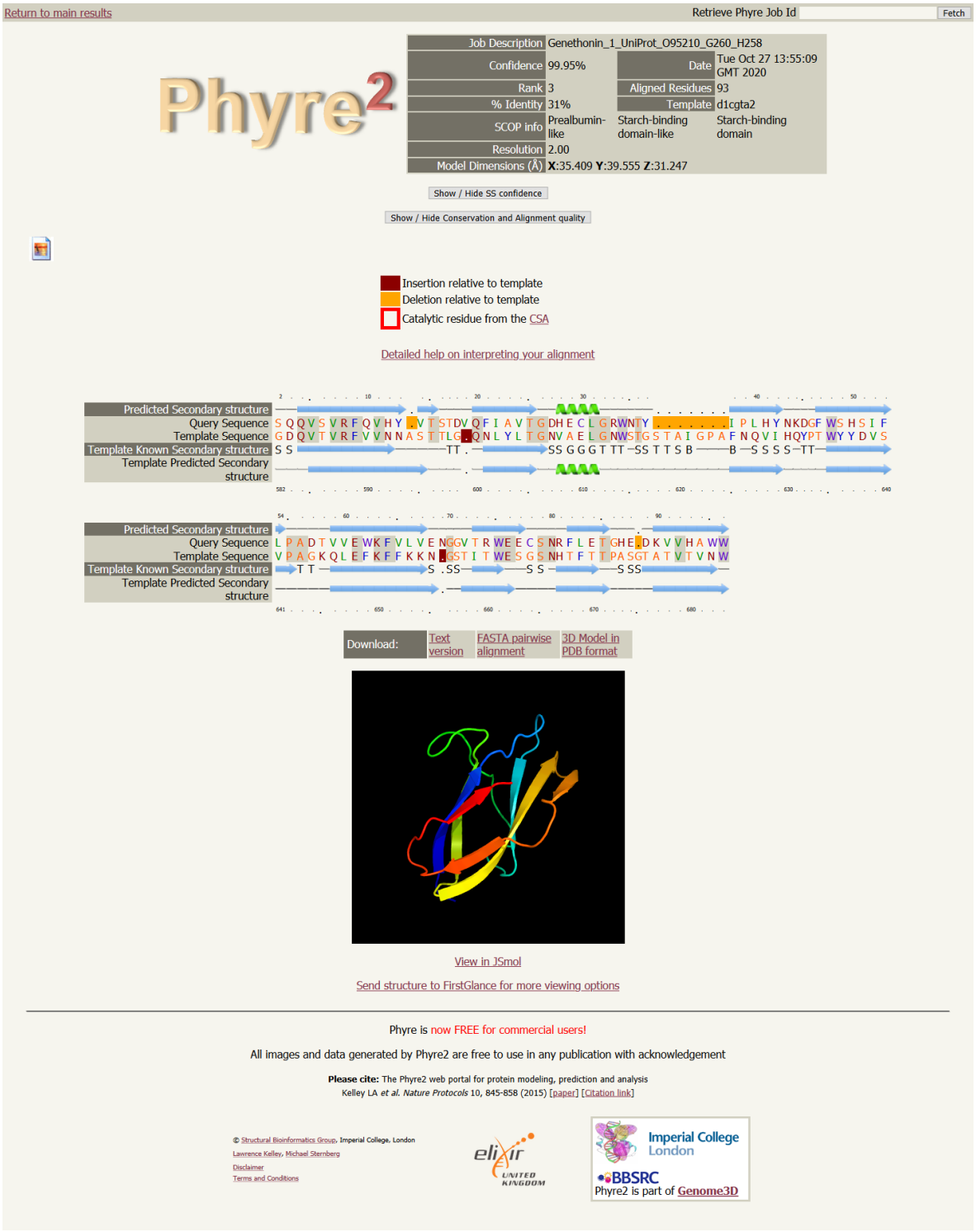
Component software
Template detection: HHpred 1.51
Secondary structure prediction: Psi-pred 2.5
Disorder prediction: Disopred 2.4
Transmembrane prediction: Memsat_SVM
Multi-template modelling and *ab initio*: Poing 1.0

Imperial College London

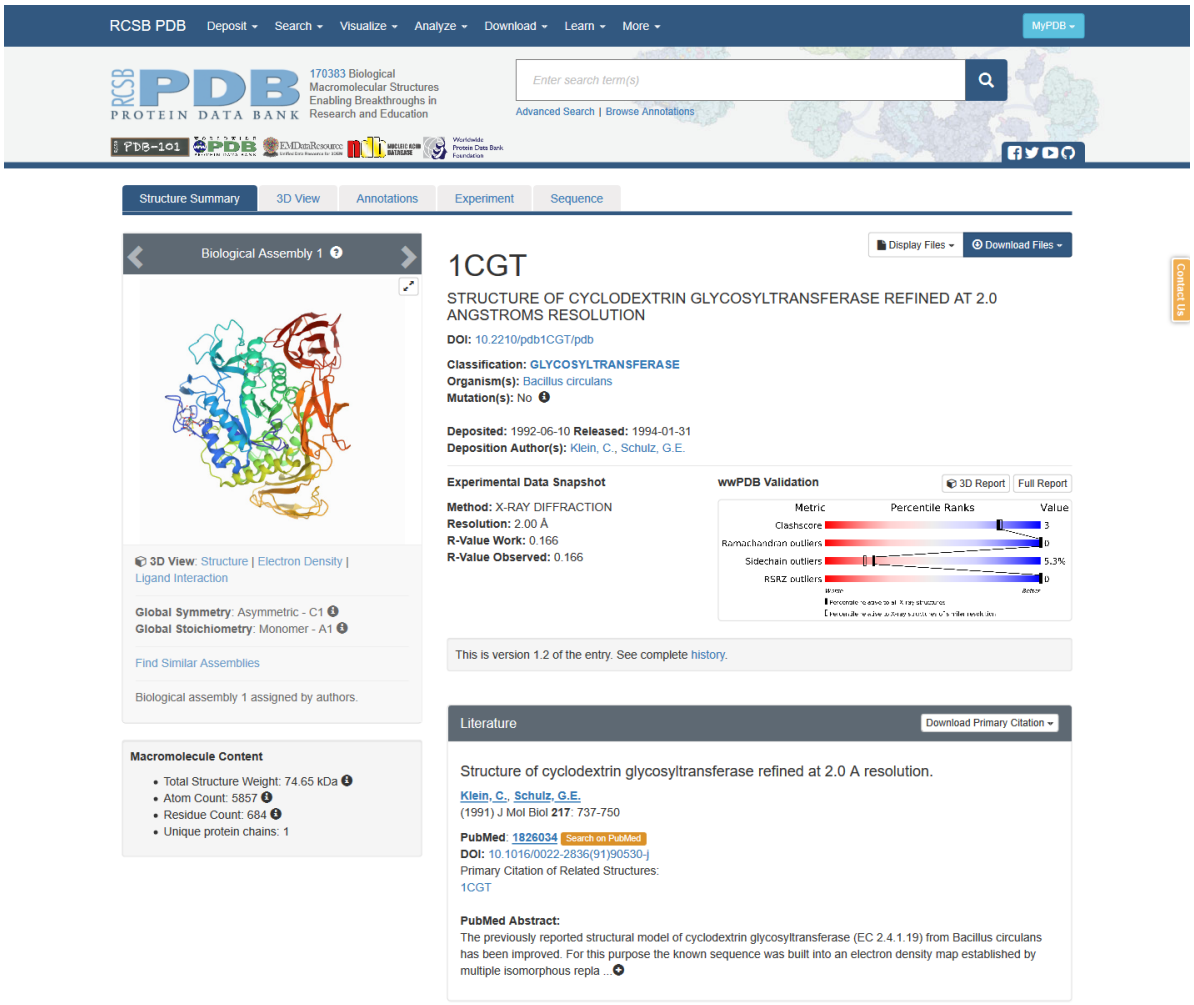
Výsledky modelovania ostávajú na serveri Phyre2 dostupné približne 1 mesiac; v každom prípade je najlepšie si ich stiahnuť zo servera a uložiť vo vlastnom počítači – na to je možné využiť tlačítko na titulnej strane hneď pod hlavičkou: „Download zip of all results“. Je veľkou výhodou, že výsledky po rozbalení sú k dizpozícii v úplne totožnej forme aj vizuálnom spracovaní tak, ako boli v origináli na serveri.

Výsledky modelovania zo serveru Phyre2:

- zobrazené sú rozkliknuté výsledky zrovnania sekvencie modelovaného proteínu („Query Sequence“) a sekvencie proteínu, ktorého štruktúra bola použitá ako templát („Template sequence“) v rámci daného sekvenčného prekryvu („Alignment Coverage“);
- toto zrovnanie je založené na štruktúre („structure-based alignment“);
- konkrétny templát: „d1cgta2“ (1CGT); úsek sekvencie: Gly582-Trp683.



PDB stránka s načítanými údajmi pre PDB kód: 1CGT (templát):



Vzájomné porovnanie terciárnych štruktúr bude ukázané na porovnaní modelu C-terminálneho úseku proteínu genetonín-1 (bol namodelovaný úsek sekvencie Ser261-Trp355 z celkového úseku Gly260-His358) s reálnou štruktúrou škrob-viažucej domény z rodiny CBM20, ktorá je prítomná v enzýme cyklodextríglukanotransferáza z *Bacillus circulans* (úsek templátu: Gly582-Trp683). Z výsledkov Phyre2 bol teda zvolený model č. 3 – „d1cgta2“ (PDB kód: 1CGT) s charakteristikou: 94% pokrytie sekvencie, 100% istota a 31% sekvenčná identita na danom prekryve. Treba povedať, že rovnako kvalitné výsledky boli poskytnuté aj pre model č. 1 – „d1cyga2“ (PDB kód: 1CYG; cyklodextríglukanotransferáza z *Bacillus stearothermophilus*). Model č. 3 však bol zvolený kvôli dostupnosti literatúry, v ktorej bola terciárna štruktúra enzýmu popísaná (PubMed ID: 1826034).

Pre samotné porovnanie štruktúr je potrebné pripraviť si koordináty aj modelu, aj templátu. Štruktúra modelu sa získa uložením z výsledkov Phyre2 serveru ako model č. 3. Štruktúra templátu sa získa z databázy PDB podľa kódu 1CGT. Tento súbor však obsahuje koordináty pre celú molekulu cyklodextríglukanotransferázy (684 zvyškov), takže pre ďalšiu prácu je výhodné si z PDB súboru vystrihnúť len koordináty pre časť štruktúry, ktorá slúžila ako templát (a ktorá v podstate predstavuje celú škrob-viažucu doménu typu CBM20 tohto enzýmu), t.j. úsek Gly582-Trp683.

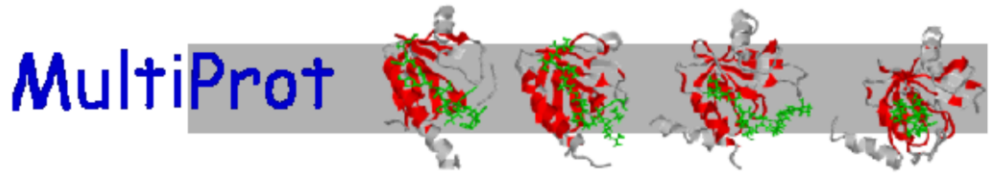
Takto boli pripravené dva súbory štruktúr v PDB formáte: (i) model č. 3 C-terminálnej oblasti ľudského genetonínu-1 – „d1cgta2_3.pdb“, t.j. úsek Ser261-Trp355 (vo výsledkoch Phyre2 je už upravené číslovanie Ser2-Trp95; pričom číslica „3“ v názve súboru naznačuje, že to je model č. 3); a (ii) štruktúra škrob-viažucej domény z cyklodextríglukanotransferázy z *Bacillus circulans* – „1CGT_CBM20.pdb“, t.j. úsek Gly582-Trp683.

Na porovnávanie terciárnych štruktúr proteínov je tiež možné využiť rôzne programy, prípadne servery; jednu z dobre etablovaných možností predstavuje server MultiProt. Pre uskutočnenie porovnania štruktúr je potrebné si súbory porovnávaných štruktúr uložiť do ZIP súboru. Po načítaní sú serverom poskytnuté výsledky vo forme zrovnania sekvencií založenom na štruktúre, pričom porovnanie je charakterizované dvomi parametrami: (i) počtom korešpondujúcich Cα atómov („Alignment Size“); a (ii) hodnotou odchýlky RMSD („root mean square deviation“) – v jednotkách „Å“, ktorá udáva mieru priemernej vzdialenosti medzi korešpondujúcimi Cα atómami. Dve porovnávané štruktúry proteínov sú si tým viac podobné, t.j. ich vzájomná štruktúrna superpozícia (prekryv; „structure overlay“; „structural superimposition“) je tým lepšia, čím vyšší je počet korešpondujúcich Cα atómov a čím menšia je hodnota odchýlky RMSD. Počet korešpondujúcich Cα atómov je vzdy menší, resp. môže byť maximálne taký veľký, ako je počet Cα atómov menšej porovnávanej molekuly proteínu.

MultiProt sever poskytuje spravidla viacero výsledkov, z ktorých je potrebné vždy zvoliť tie, ktoré spĺňajú kritéria o čo najväčšom možnom počte korešpondujúcich Cα atómov a čo najnižšej hodnote odchýlky RMSD. Pre prípad štruktúrneho porovnania modelu C-terminálnej oblasti genetonínu-1

a reálnej štruktúry škrob-viažucej domény z rodiny CBM20 prítomnej v cyklodextríglukanotransferáze sú to výsledky v 1. riadku, t.j. „Alignment Size“: 93 a „RMSD“: 0,59 Å. Počet korešpondujúcich Cα atómov v porovnávaných proteínových molekulách bol 95 pre C-terminálny región genetonínu-1 (súbor: „d1cgta2_3.pdb“) a 102 pre škrob-viažucu doménu (súbor: „1CGT_CBM20.pdb“). Pre uvedené výsledky superpozície (prekryvu) je nakoniec potrebné stiahnuť zo serveru súbor „aligned.pdb“, v ktorom sú uložené koordináty oboch proteínových molekúl po ich vzájomnej superpozícii a ktorý je možné vo vhodnom programe zobraziť (obr. 6.3).

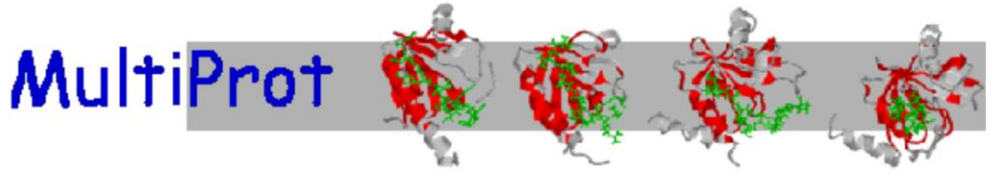
Úvodná web-stránka serveru MultiProt:
http://bioinfo3d.cs.tau.ac.il/MultiProt/



[Back to Home Page]

PDB IDs separated by a space: *(e.g. 1adj:A 1hc7:A 1v95)*
Upload zip file of pdb structures: Prehľadávať... Nie je zvolený súbor.
Accuracy (in Angstroms): 3
Sequence Order: Yes ◉ No ○
Odoslať Clear

Výsledky zo serveru MultiProt:

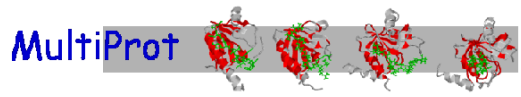


Protein	Number of C-alpha atoms
1CGT_CBM20	102
d1cgta2_3	95

Number of aligned molecules : 2

Alignment Size	RMSD	Molecules	PDB alignment
93	0.59	1CGT_CBM20 d1cgta2_3	aligned.pdb
15	1.84	1CGT_CBM20 d1cgta2_3	aligned.pdb
15	2.11	1CGT_CBM20 d1cgta2_3	aligned.pdb

Zovnanie sekvencií generované serverom MultiProt:

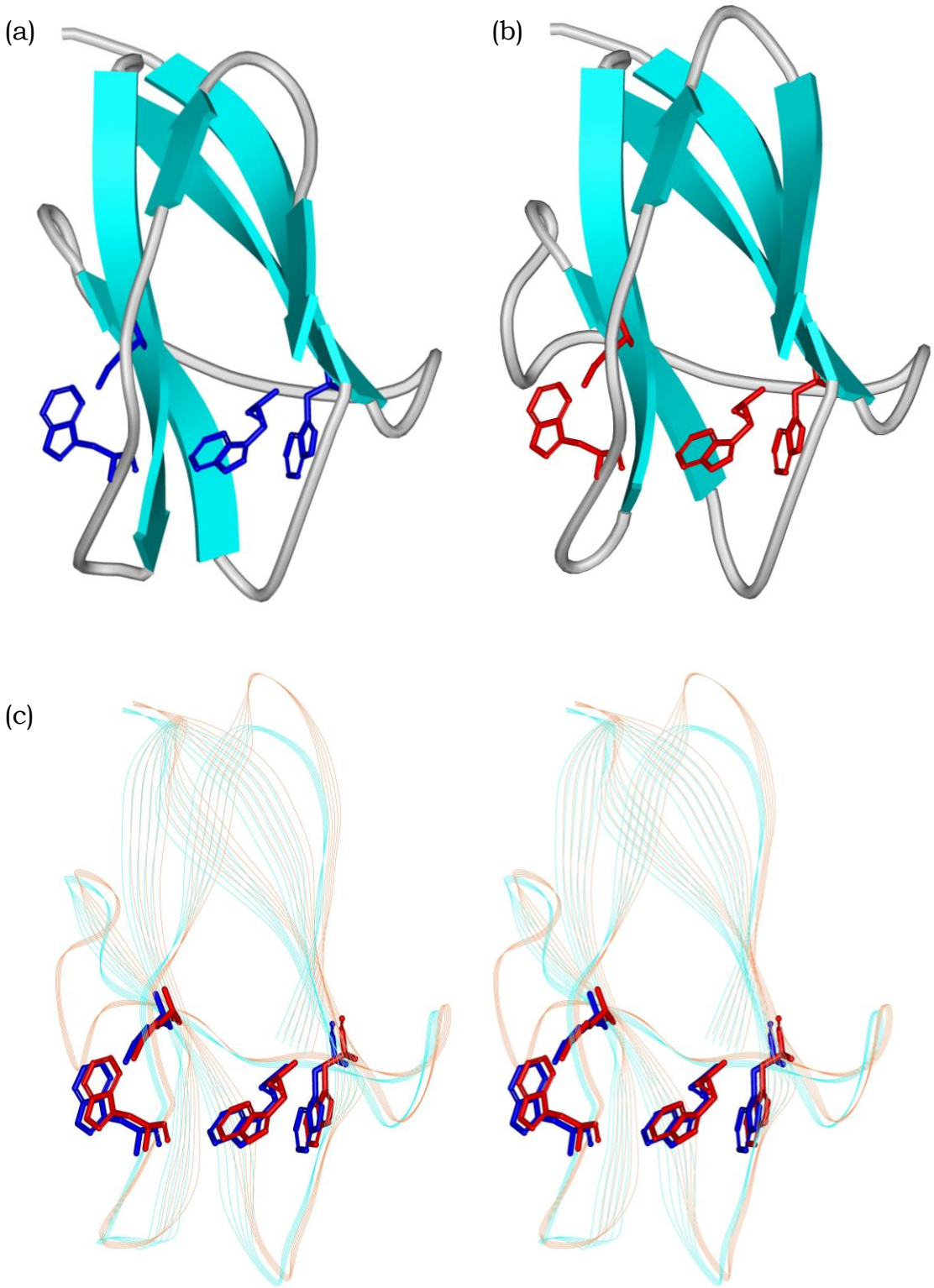


Number of aligned molecules : 2

Alignment Size	RMSD	Molecules	PDB alignment
93	0.59	1CGT_CBM20 d1cgta2_3	aligned.pdb

Alignment syntaxis: *ChainID.AminoAcidCode.ResidueNumber*

1CGT_CBM20	d1cgta2_3
A G 582	S 2
A D 583	Q 3
A Q 584	Q 4
A V 585	V 5
A T 586	S 6
A V 587	V 7
A R 588	R 8
A F 589	F 9
A V 590	Q 10
A V 591	V 11
A N 592	H 12
A N 593	Y 13
A A 594	V 14
A S 595	T 15
A T 596	S 16
A T 597	T 17
A L 598	D 18
A G 599	V 19
A Q 600	Q 20
A N 601	F 21
A L 602	I 22
A Y 603	A 23
A L 604	V 24
A T 605	T 25
A G 606	G 26
A N 607	D 27
A V 608	H 28
A A 609	E 29
A E 610	C 30
A L 611	L 31
A G 612	G 32
A A 620	R 33
A I 621	W 34
A G 622	N 35
A P 623	T 36
A A 624	Y 37
A F 625	I 38
A N 626	P 39
A Q 627	L 40
A V 628	H 41
A I 629	Y 42
A H 630	N 43
A Q 631	K 44
A Y 632	D 45
A P 633	G 46
A T 634	F 47
A W 635	W 48
A Y 636	S 49
A Y 637	H 50
A D 638	S 51
A V 639	I 52
A S 640	F 53
A V 641	L 54
A P 642	P 55
A A 643	A 56
A G 644	D 57
A K 645	T 58
A Q 646	V 59
A L 647	V 60
A E 648	E 61
A F 649	W 62
A K 650	K 63
A F 651	F 64
A F 652	V 65
A K 653	L 66
A K 654	V 67
A N 655	E 68
A G 656	G 70
A S 657	G 71
A T 658	V 72
A I 659	T 73
A T 660	R 74
A W 661	W 75
A E 662	E 76
A S 663	E 77
A G 664	C 78
A S 665	S 79
A N 666	N 80
A H 667	R 81
A T 668	F 82
A F 669	L 83
A T 670	E 84
A T 671	T 85
A P 672	G 86
A G 675	E 88
A T 676	D 89
A A 677	K 90
A T 678	V 91
A V 679	V 92
A T 680	H 93
A V 681	A 94
A N 682	W 95
A W 683	W 96



Obr. 6.3. Znázornenie (a) modelu C-terminálnej oblasti ľudského genetonínu-1 a (b) terciárnej štruktúry škrob-viažucej domény z rodiny CBM20 cyklodextrínglukanotransferázy z *Bacillus circulans*. Elementy sekundárnej štruktúry – β-vlákna – sú zobrazené ako tyrkysové šípky. (c) Štruktúrny prekryv (znázornený stereo; polypeptidový reťazec ako líniová stuha) modelu potenciálnej škrob-viažucej domény z genetonínu-1 (tyrkysová) a reálnej škrob-viažucej domény z cyklodextrínglukanotransferázy (oranžová). Selektované zvyšky (modré pre genetonín-1; červené pre cyklodextrínglukanotransferázu) sú dôležité pre funkciu väzby α-glukánov pre škrob-viažuce domény z rodiny CBM20. Zobrazené v programe WebLabViewerLite (Molecular Simulations, Inc.).

7. Praktické úlohy a cvičenia

7.1. Sledovanie veľkosti genómov rôznych mikroorganizmov

Zadanie

Na základe PubMed ID čísiel pre jednotlivé literárne záznamy, zistite a navzájom porovnajte veľkosti kompletných genómov vybraných mikroorganizmov, ako aj doplňte ďalšie údaje do tab. 7.1:

- časopis, v ktorom bola štúdia uverejnená;
- rok vydania;
- doménu prokaryotov – buď *Bacteria* alebo *Archaea*;
- veľkosť genómu v bázových pároch;
- počet génov kódujúcich proteíny (ORF).

Tabuľka 7.1. Príklady organizmov s kompletne sekvenovaným genómom.

Č.	Mikroorganizmus	PubMed	Časopis	Rok	Doména	Veľkosť genómu	ORF
1	*Haemophilus influenzae*	7542800					
2	*Mycoplasma genitalium*	7569993					
3	*Methanococcus jannaschii*	8688087					
4	*Helicobacter pylori*	9252185					
5	*Escherichia coli*	9278503					
6	*Bacillus subtilis*	9384377					
7	*Archaeoglobus fulgidus*	9389475					
8	*Borrelia burgdorferi*	9403685					
9	*Mycobacterium tuberculosis*	9634230					
10	*Treponema pallidum*	9665876					
11	*Thermotoga maritima*	10360571					
12	*Deinococcus radiodurans*	10567266					
13	*Vibrio cholerae*	10952301					
14	*Yersinia pestis*	11586360					
15	*Shewanella oneidensis*	12368813					

7.2. *In silico* analýza triózafosfátizomerázy

Zadanie

Vykonajte bioinformatickú analýzu enzýmu triózafosfátizomeráza pochádzajúceho zo 40 rôznych zdrojov zahŕňajúcich baktérie, archeóny a eukaryoty podľa tab. 7.2.

(1) Vytvorte si priečinok „TIM“.

(2) Všetkých 40 sekvencií TIM zhromaždite z databázy UniProt do vstupného súboru („TIM.txt“; textový súbor) vhodného pre program Clustal-Omega.

(3) Na EBI serveri zrovnajte sekvencie TIM v programe Clustal-Omega – http://www.ebi.ac.uk/Tools/msa/clustalo/ – získajte dva súbory: „TIM_aln.txt“ (alignment „with character counts“) a „TIM_fas.txt“ (formát Pearson/FASTA); dodržte vstupné poradie sekvencií (input order).

(4) Súbor „TIM_aln.txt“ otvorte v programe MS-Word a v zrovnaných sekvenciách zvýraznite žltým podfarbením aminokyselinové zvyšky aktívneho miesta (Asn10, Lys12, His95, Glu165 v sekvencii triózafosfátizomerázy zo *Saccharomyces cerevisiae*); súbor uložte ako dokument „TIM_aln.doc“.

(5) Vypočítajte hodnoty CL, SI a SS.

(6) Na EBI serveri v rámci programu Simple Phylogeny – http://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/ – vypočítajte dva evolučné stromy pre TIM – s uvažovaním medzier v sekvenciách („TIM_off.txt“) a s ich ignorovaním („TIM_on.txt“) – na základe finálneho súboru „TIM_fas.txt“.

(7) Vypočítané evolučné stromy zobrazte v programe iTOL (interactive Tree of Life; https://itol.embl.de/), vložte ako exportované obrázky (napr. PNG) do súboru „TIM_aln.doc“, porovnajte a zapíšte diskusiu celej práce.

Tabuľka 7.2. Zoznam študovaných triózafosfátizomeráz.

Č.	Zdroj	Skratka	UniProt	Dĺžka
	Bacteria			
1	*Agrobacterium tumefaciens*	Agrtu	Q8UEY3	256
2	*Nostoc* sp. PCC 7120	Nossp	Q8YP17	241
3	*Bacillus anthracis*	Bacan	Q81X76	251
4	*Bifidobacterium longum*	Biflo	Q8G6D5	267
5	*Borrelia burgdorferi*	Borbu	Q59182	253
6	*Chlamydia pneumoniae*	Chlpn	Q9Z6J6	254
7	*Deinococcus radiodurans*	Deira	Q9RUP5	244
8	*Escherichia coli*	Escco	P0A858	255
9	*Haemophilus influenzae*	Haein	P43727	263
10	*Mycobacterium tuberculosis*	Myctu	P9WG42	261
11	*Mycoplasma genitalium*	Mypge	P47670	244
12	*Neisseria meningitidis*	Neime	Q9JW31	251
13	*Salmonella typhimurium*	Salty	Q8ZKP7	*255*
14	*Staphylococcus aureus*	Staau	Q5HHP3	253
15	*Streptococcus mutans*	Stcmu	P72484	252
16	*Streptomyces coelicolor*	Strco	Q9Z520	258
17	*Synechocystis* sp. PCC 6803	Synsp	Q59994	242
18	*Thermotoga maritima*	Thtma	P36204	255 (400-654)
19	*Vibrio cholerae*	Vibch	Q9KNR1	257
20	*Yersinia pestis*	Yerpe	Q8ZJK9	255
	Archaea			
21	*Aeropyrum pernix*	Aerpe	Q9YBR1	223
22	*Archaeoglobus fulgidus*	Arcfu	O28965	223
23	*Haloarcula marismortui*	Haama	Q5V4J7	215
24	*Halobacterium salinarium*	Habsa	Q9HQS4	214
25	*Methanopyrus kandleri*	Mepka	Q8TUT9	226
26	*Methanosarcina mazei*	Mesma	Q8PXE2	222
27	*Pyrobaculum aerophilum*	Pybae	Q8ZX28	227
28	*Pyrococcus furiosus*	Pycfu	P62002	228
29	*Sulfolobus solfataricus*	Sulso	Q97VM8	227
30	*Thermoproteus tenax*	Thpte	Q8NKN9	226
	Eucarya			
31	*Arabidopsis thaliana* (mouse-ear cress)	Arath	Q9SKP6	255 (61-315)
32	*Aspergillus nidulans* (fungi)	Aspni	P04828	249
33	*Caenorhabditis elegans* (nematode)	Caeel	Q10657	247
34	*Drosophila melanogaster* (fruit fly)	Drome	P29613	247
35	*Gracilaria verrucosa* (red alga)	Grave	P48492	250
36	*Homo sapiens* (human; isoform 1)	Homsa	P60174	249
37	*Oryctolagus cuniculus* (rabbit, isoform 1)	Orlcu	P00939	249
38	*Oryza sativa* (rice)	Orysa	P48494	253
39	*Pan troglodytes* (chimpanzee)	Pantr	P60175	249
40	*Saccharomyces cerevisiae* (yeast)	Sacce	P00942	248

7.3. Bioinformatická analýza pektolytických enzýmov z rodiny GH28

Zadanie

Vykonajte bioinformatickú analýzu pektolytických enzýmov z rodiny GH28 pochádzajúcich z 25 rôznych zdrojov zahŕňajúcich baktérie a eukaryoty podľa tab. 7.3.

(1) Vytvorte si priečinok „GH28“.

(2) Všetkých 25 sekvencií pektolytických enzýmov zhromaždite z databázy UniProt do vstupného súboru („GH28.txt“; textový súbor) vhodného pre program Clustal-Omega.

(3) Na EBI serveri zrovnajte sekvencie enzýmov z rodiny GH28 v programe Clustal-Omega – http://www.ebi.ac.uk/Tools/msa/clustalo/ – získajte dva súbory: „GH28_aln.txt“ (alignment „with character counts“) a „GH28_fas.txt“ (formát Pearson/FASTA); dodržte vstupné poradie sekvencií (input order).

(4) Súbor „GH28_aln.txt“ otvorte v programe MS-Word a v zrovnaných sekvenciách zvýraznite žltým podfarbením konzervované sekvenčné regióny (udané pre tri reprezentatívne sekvencie nižšie); súbor uložte ako dokument „GH28_aln.doc“.

Konzervované úseky vybraných členov rodiny GH28:

Proteín č. 4:	**Aspni_PG**	**178_NTD**	**201_DD**	**222_GHG**	**256_RIK**	**Y291**
Proteín č. 21:	**Asptu_XGH**	**205_NTD**	**228_DD**	**250_SHG**	**284_GIK**	**Y322**
Proteín č. 22:	**Aspac_RG**	**193_GLD**	**215_DE**	**237_SGG**	**269_MIK**	**W302**

(5) Vypočítajte hodnoty CL, SI a SS.

(6) Na EBI serveri v rámci programu Simple Phylogeny – http://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/ – vypočítajte dva evolučné stromy pre enzýmy z rodiny GH28 – s uvažovaním medzier v sekvenciách („GH28_off.txt“) a s ich ignorovaním („GH28_on.txt“) – na základe finálneho súboru „GH28_fas.txt“.

(7) Vypočítané evolučné stromy zobrazte v programe iTOL (interactive Tree of Life; https://itol.embl.de/), vložte ako exportované obrázky (napr. PNG) do súboru „GH28_aln.doc“, porovnajte a zapíšte diskusiu celej práce.

Tabuľka 7.3. Zoznam študovaných pektolytických enzýmov z rodiny GH28.

Č.	Enzým	Zdroj	Ríša	Skratka	EC	UniProt
1	Polygalakturonáza	*Erwinia carotovora*	Baktérie	Erwca_PG	3.2.1.15	P26509
2	Polygalakturonáza	*Ralstonia solanacearum*	Baktérie	Ralso_PG	3.2.1.15	P20041
3	Polygalakturonáza	*Thermotoga maritima*	Baktérie	Thema_PG	3.2.1.15	Q9WYR8
4	Polygalakturonáza	*Aspergillus niger*	Huby	Aspni_PG	3.2.1.15	P26214
5	Polygalakturonáza	*Aspergillus parasiticus*	Huby	Asppa_PG	3.2.1.15	P49575
6	Polygalakturonáza	*Claviceps purpurea*	Huby	Clapu_PG	3.2.1.15	P78607
7	Polygalakturonáza	*Fusarium moniliforme*	Huby	Fusmo_PG	3.2.1.15	Q07181
8	Polygalakturonáza	*Kluyveromyces marxianus*	Huby	Kluma_PG	3.2.1.15	O13478
9	Polygalakturonáza	*Penicillium expansum*	Huby	Penex_PG	3.2.1.15	O59925
10	Polygalakturonáza	*Saccharomyces cerevisiae*	Huby	Sacce_PG	3.2.1.15	P47180
11	Polygalakturonáza	*Actinidia deliciosa*	Rastliny	Actde_PG	3.2.1.15	P35336
12	Polygalakturonáza	*Lycopersicon esculentum*	Rastliny	Lyces_PG	3.2.1.15	P05117
13	Polygalakturonáza	*Malus domestica*	Rastliny	Maldo_PG	3.2.1.15	P48978
14	Polygalakturonáza	*Phaedon cochleariae*	Hmyz	Phaco_PG	3.2.1.15	O97400
15	Exopolygalakturonáza	*Erwinia chrysanthemi*	Baktérie	Erwch_EPG	3.2.1.82	P15922
16	Exopolygalakturonáza	*Ralstonia solanacearum*	Baktérie	Ralso_EPG	3.2.1.82	Q53241
17	Exopolygalakturonáza	*Yersinia enterocolitica*	Baktérie	Yeren_EPG	3.2.1.82	O68975
18	Exopolygalakturonáza	*Aspergillus tubigensis*	Huby	Asptu_EPG	3.2.1.67	Q00293
19	Exopolygalakturonáza	*Fusarium oxysporum*	Huby	Fusox_EPG	3.2.1.67	O74255
20	Peľová polygalakturonáza	*Zea mays*	Rastliny	Zeama_PPG	3.2.1.67	P26216
21	Xylogalakturonanhydroláza	*Aspergillus tubigensis*	Huby	Asptu_XGH	3.2.1.-	Q9UUZ2
22	Ramnogalakturonáza	*Aspergillus aculeatus*	Huby	Aspac_RG	3.2.1.-	Q00001
23	Ramnogalakturonáza A	*Aspergillus niger*	Huby	AspniA_RG	3.2.1.-	P87160
24	Ramnogalakturonáza B	*Aspergillus niger*	Huby	AspniB_RG	3.2.1.-	P87161
25	Ramnogalakturonáza	*Botryotinia fuckeliana*	Huby	Botfu_RG	3.2.1.-	P87247

7.4. Bioinformatická analýza α-amyláz z rodiny GH13

Zadanie

Vykonajte bioinformatickú analýzu α-amyláz z rodiny GH13 pochádzajúcich z rôznych zdrojov, ktoré na základe sekvenčnej podobnosti patria do viac ako 10 podrodín tejto rodiny podľa tab. 7.4.

(1) Vytvorte si priečinok „GH13“.

(2) Všetkých 28 sekvencií α-amyláz zhromaždite z databázy UniProt do vstupného súboru („GH13.txt“; textový súbor) vhodného pre program Clustal-Omega.

(3) Na EBI serveri zrovnajte sekvencie enzýmov z rodiny GH13 v programe Clustal-Omega – http://www.ebi.ac.uk/Tools/msa/clustalo/ – získajte dva súbory: „GH13_aln.txt“ (alignment „with character counts“) a „GH13_fas.txt“ (formát Pearson/FASTA); dodržte vstupné poradie sekvencií (input order).

(4) Súbor „GH13_aln.txt“ otvorte v programe MS-Word a v zrovnaných sekvenciách zvýraznite tri konzervované sekvenčné regióny (KSR) okolo katalytických zvyškov: Asp227 (bez signálneho peptidu – 21 zvyškov: 206), Glu251 (230) a Asp318 (297) (počítanie v α-amyláze z *Aspergillus oryzae*; regióny sú: GLRI***D***TVKH, ***E***VLD a FVENH***D***). Zároveň skontrolujte, či sú sekvencie všetkých α-amyláz správne zrovnané, t.j. či pre každú z 28 sekvencií boli identifikované všetky tri katalytické zvyšky (v selektovaných KSR). Tiež rovnako zvýraznite ďalšie štyri KSR: 56_***G***FTAIWITP (začiatok regiónu v sekvencii α-amylázy z *Aspergillus oryzae* s uvažovaním signálneho peptidu: Gly77), 117_***D***VVANH (Asp138), 173_***L***PDLD (Leu194) a 323_***G***IPIIYAGQ (Gly344); súbor uložte ako dokument „GH13_aln.doc“. Ak niektoré KSR nie sú správne zrovnané, je potrebné zrovnanie sekvencií manuálne upraviť a tieto zmeny následne preniesť do zrovnania vo formáte Pearson/FASTA (súbor „GH13_fas.txt“), ktorý bude slúžiť na výpočet evolučného stromu.

(5) Vypočítajte hodnoty CL, SI a SS – z finálneho zrovnania (po prípadnej manuálnej úprave).

(6) Na EBI serveri v rámci programu Simple Phylogeny – http://www.ebi.ac.uk/Tools/phylogeny/simple_phylogeny/ – vypočítajte dva evolučné stromy pre enzýmy z rodiny GH13 – s uvažovaním medzier v sekvenciách („GH13_off.txt“) a s ich ignorovaním („GH13_on.txt“) – na základe finálneho súboru „GH13_fas.txt“.

(7) Vypočítané evolučné stromy zobrazte v programe iTOL (interactive Tree of Life; https://itol.embl.de/), vložte ako exportované obrázky (napr. PNG) do súboru „GH13_aln.doc“, porovnajte a zapíšte diskusiu celej práce.

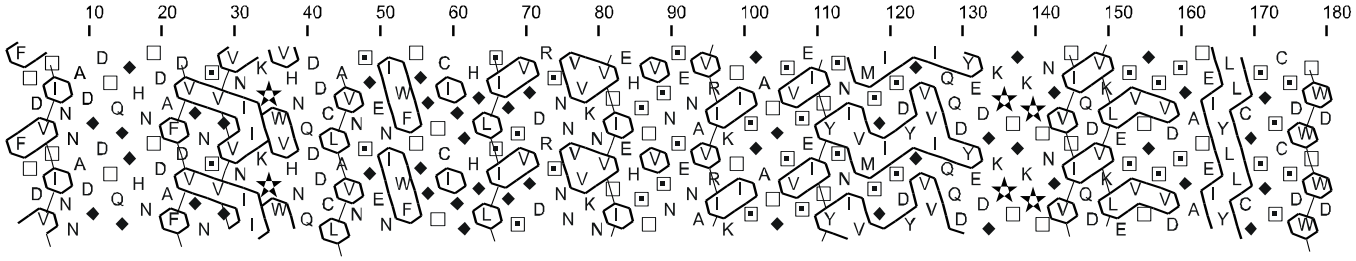
Tabuľka 7.4. Zoznam študovaných α-amyláz z rodiny GH13.

Č.	Organizmus	UniProt	Skratka	Dĺžka	Podrodina
1	*Aspergillus oryzae*	P0C1B3	Aspor_1	499	13_1
2	*Saccharomycopsis fibuligera*	D4P4Y7	Sacfi_1	494	13_1
3	*Bacillus amyloliquefaciens*	P00692	Bacam_5	514	13_5
4	*Halothermothrix orenii*	B8CZ54	Halor_5	623	13_5
5	*Hordeum vulgare*	P00693	Horvu_6	438	13_6
6	*Oryza sativa*	P17654	Orysa_6	434	13_6
7	*Pyrococcus woesei*	Q7LYT7	Pyrwo_7	460	13_7
8	*Thermococcus hydrothermalis*	O93647	Thehy_7	457	13_7
9	*Drosophila melanogaster*	P08144	Drome_15	494	13_15
10	*Tenebrio molitor*	P56634	Tenmo_15	471	13_15
11	*Gallus gallus*	Q98942	Galga_24	512	13_24
12	*Homo sapiens* (sliny)	P04745	Homsa_24	511	13_24
13	*Aeromonas hydrophila*	P22630	Aerhy_27	464	13_27
14	*Xanthomonas campestris*	Q56791	Xanca_27	475	13_27
15	*Bacillus subtilis*	Q45520	Bacsu_28	477	13_28
16	*Lactobacillus amylovorus*	Q48502	Lacam_28	953	13_28
17	*Pseudoalteromonas haloplanktis*	P29957	Pseha_32	669	13_32
18	*Streptomyces limosus*	P09794	Strli_32	566	13_32
19	*Halothermothrix orenii*	Q8GPL8	Halor_36	515	13_36
20	*Dictyoglomus thermophilum*	P14899	Dicth_36	499	13_36
21	*Uncultured bacterium*	D9MZ14	Uncba_37	639	13_37
22	*Photobacterium profundum*	Q6LIA8	Phopr_37	687	13_37
23	*Roseburia* sp. A2-194	Q3LB10	Rossp_41	1674	13_41
24	*Micrococcus* sp. 207	Q06812	Micsp_41	1104	13_41
25	*Bacillus aquimaris*	G8IJA7	Bacaq_xx	512	xx
26	*Geobacillus thermoleovorans*	G8N704	Geoth_xx	511	xx
27	*Haloarcula hispanica*	Q4A3E0	Halhi_yy	433	yy
28	*Natronoarchaeum philippinense*	A0A285N7G2	Natph_yy	511	yy

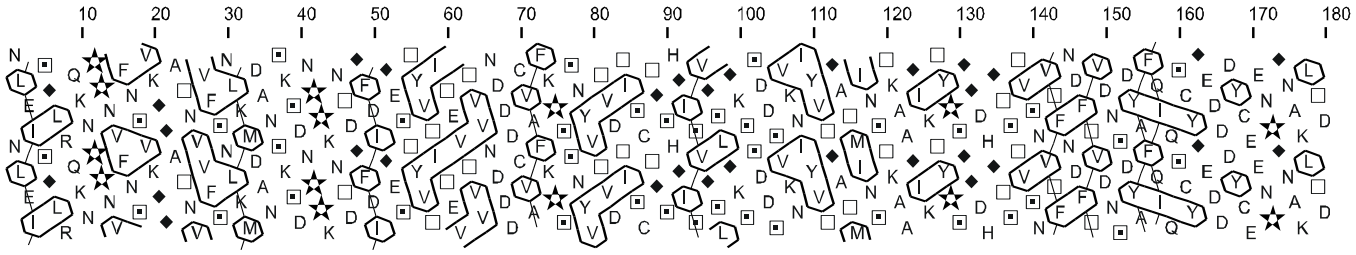
7.5. Analýza enzýmov z rodiny GH28 pomocou metódy HCA

V sekvencii polygalakturonázy z *Aspergillus niger* (Aspni_PG) identifikujte úseky 18_NTD, 41_DD, 62_GHG, 96_RIK a zvyšok Tyr131. Potom k nim nájdite korešpondujúce úseky v sekvenciách xylogalakturonanhydrolázy z *Aspergillus tubigensis* (Asptu_XGH) a ramnogalakturonázy z *Aspergillus aculeatus* (Aspac_RG). Výsledky dajte do súvisu s výsledkami *in silico* analýzy pektolytických enzýmov z rodiny GH28, najmä s ohľadom na ich konzervované úseky (zadanie č. 7.3, bod č. 4).

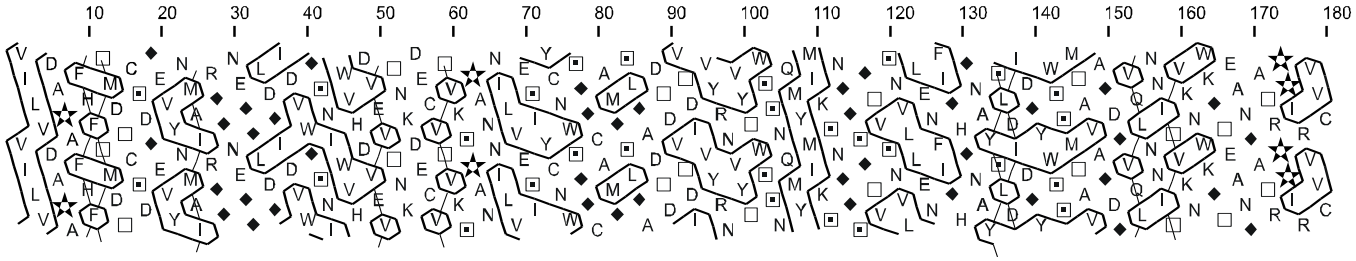
Aspni_PG



Asptu_XGH



Aspac_RG



7.6. Analýza hémových kataláz pomocou metódy HCA

Teória

Hémové peroxidázy sú oxidoreduktázy odpovedajúce na rôzne formy oxidačného stresu. Trieda *Class I* hémových peroxidáz obsahuje tri skupiny enzýmov: (i) katalázy-peroxidázy (CP); (ii) askorbát-peroxidázy (APX); a (iii) cytochróm c peroxidázy (CCP). Kataláza-peroxidáza (CP) je jediným enzýmom skupiny, ktorý má *katalázovú* aktivitu, t.j. je schopný aj oxidovať, aj redukovať H_2O_2. Ostatné *nekatalázové* enzýmy (APX a CCP) dokážu peroxid vodíka len redukovať. K funkcii hémových peroxidáz je potrebná prostetická skupina – ióny Fe vo forme hému. Hém je viazaný dvomi His zvyškami: tzv. vzdialená strana väzbového zoskupenia – bližšie k N-koncu proteínu (His87 v CP z *Archaeoglobus fulgidus*) a tzv. blízka strana väzbového zoskupenia – ďalej od N-konca proteínu (His249 v CP z *Archaeoglobus fulgidus*). V sekvencii všetkých členov triedy *Class I* hémových peroxidáz sú tri konzervované regióny, ktoré obsahujú funkčne esenciálne aminokyselinové zvyšky (v CP z *Archaeoglobus fulgidus*): (i) His87 (vzdialená strana): 79_PLFIRLAW***H***SAGSYR_93; (ii) His249 (blízka strana): 242_VALIAGG***H***AFGKC_254; a (iii) Asp359 (vodíková väzba s His249): 353_PRMLTA***D***LALRF_364. CP obsahujú v porovnaní s APX a CCP duplikát celej katalytickej domény (sú zhruba 2x dlhšie). Originál, t.j. katalyticky aktívna časť proteínu, sa nachádza v N-terminálnej časti, kým duplikát, ktorý je katalyticky neaktívny, sa nachádza v C-terminálnej oblasti molekuly. Duplikát je katalyticky neaktívny preto, že aj keď obsahuje všetky tri konzervované regióny, His ligandy sú substituované inými zvyškami.

Zadanie

(1) Vytvorte na ploche Vášho počítača priečinok s názvom „HCA“.

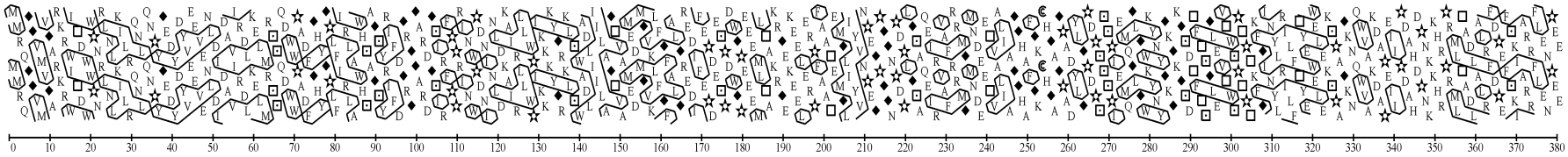
(2) Z databázy UniProt získajte sekvencie enzýmov: (i) kataláza-peroxidáza z *Archaeoglobus fulgidus* (UniProt: O28050); (ii) askorbát-peroxidáza z *Arabidopsis thaliana* (cytosolic; Q05431); (iii) askorbát-peroxidáza z *Arabidopsis thaliana* (thylakoid-bound; Q42593); a (iv) cytochróm c peroxidáza zo *Saccharomyces cerevisiae* (P00431).

(3) Pomocou aktuálneho HCA-servera vytvorte na internete HCA obrazy všetkých štyroch sekvencií hémových peroxidáz a uložte ich do priečinku „HCA“ vo formáte PDF.

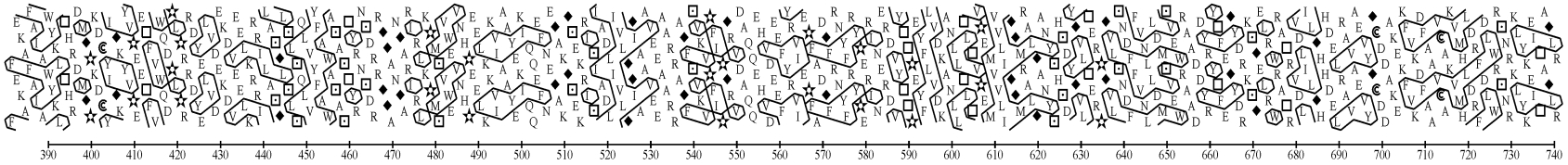
(4) Na vytlačenom HCA obrázku študovaných hémových peroxidáz identifikujte a farebne vyznačte: (i) tri konzervované sekvenčné regióny (KSR) obsahujúce His87, His249 a Asp359 v HCA obraze CP z *A. fulgidus* (Arcfu_CP.txt); (ii) tri KSR (t.j. oblasti korešpondujúce His87, His249 a Asp359) v oboch APX a CCP (Arath_APX, Arath_APXT a Sacce_CCP); a (iii) oblasti korešpondujúce His87, His249 a Asp359 v C-terminálnom duplikáte katalytickej domény CP z *A. fulgidus*.

Arcfu_CP

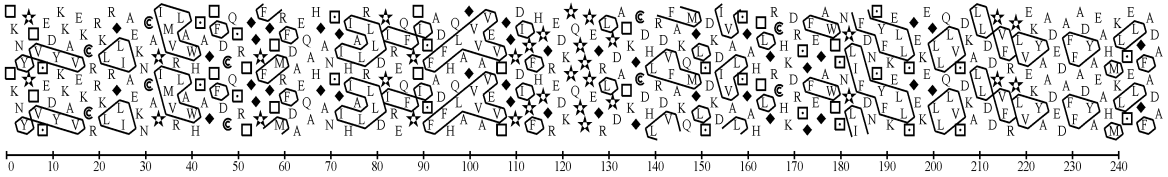
N-terminána kópia



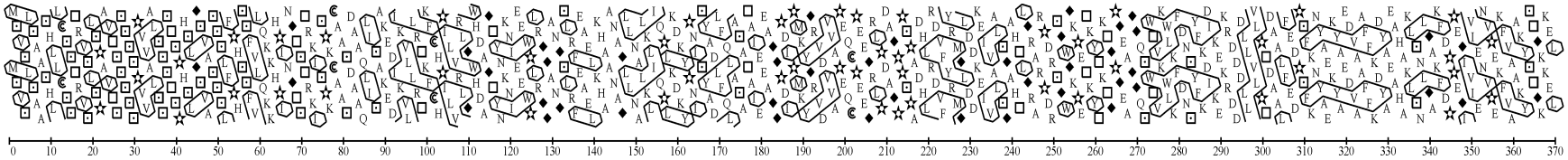
C-terminálna kópia



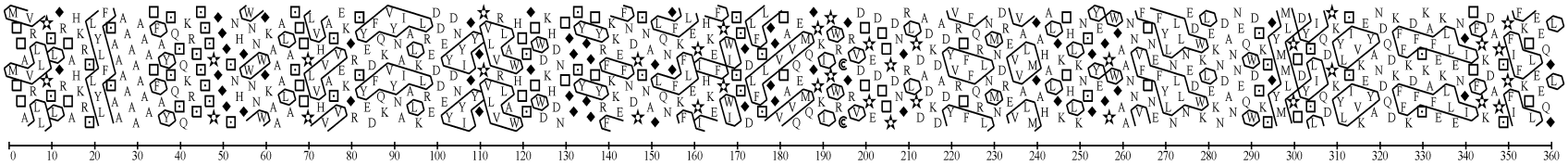
Arath_APX



Arath_APXT



Sacce_CCP



7.7. Analýza škrob-viažucich domén pomocou metódy HCA

Teória

Škrob-viažuca doména (starch-binding domain; SBD) je klasifikovaná ako tzv. modul z rodiny CBM20, t.j. „carbohydrate-binding module family 20“. Táto doména umožňuje amylolytickému enzýmu viazať a degradovať surový (nezmazovatený) škrob. CBM20 sa môže nachádzať v α-amylázach, β-amylázach a glukoamylázach z baktérií, archeónov, kvasiniek a vláknitých húb, ktorých katalytické domény sú navzájom sekvenčne nepodobné; ale napr. aj v rastlinných 4-α-glukanotransferázach. U živočíchov môže byť súčasťou glukánfosfatázy laforín, prípadne aj doteraz bližšie necharakterizovaného proteínu genetonín-1, pričom u oboch funguje skôr ako glykogén-viažuca doména (glycogen-binding domain; GBD). Sekvenčne príbuzná doména SBD/GBD typu CBM48 sa môže nachádzať aj v ďalších regulačných proteínoch, ktoré sú zapojené do regulácie metabolizmu škrobu u rastlín (proteín-4 nadbytku škrobu; SEX-4) a glykogénu u živočíchov (súčasť β-podjednotky AMP-aktivovanej proteínovej kinázy; AMPKβ).

Zadanie

Na HCA zobrazení vyššie spomenutých enzýmov a proteínov identifikujte ich škrob-, resp. glykogén-viažucu doménu. Vo všeobecnosti má doména vo všetkých prípadoch približne 100 aminokyselinových zvyškov, pričom sa môže nachádzať ako na N-, tak aj na C-konci proteínovej molekuly, prípadne aj vo vnútri proteínového reťazca. Pri identifikácii si pomôžte nasledovnými charakteristickými sekvenčnými črtami SBD typu CBM20 z amylolytických enzýmov:

N-koniec domény

~ 30 zvyškov

LG-W

~15 zvyškov

P-W

~ 15 zvyškov

K

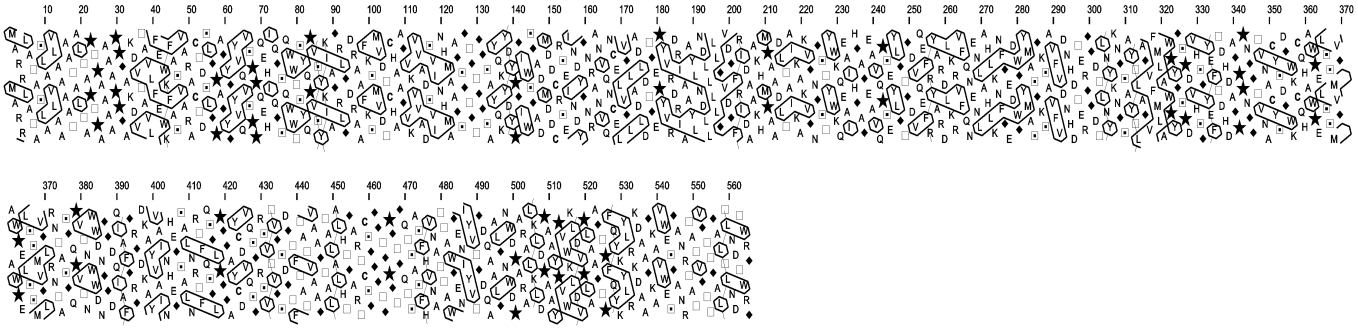
~ 10 zvyškov

W----N

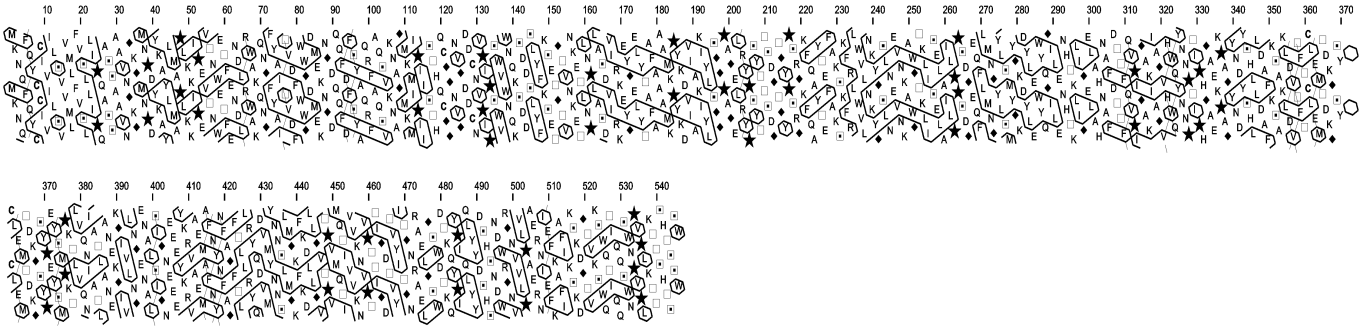
~ 15 zvyškov

C-koniec domény

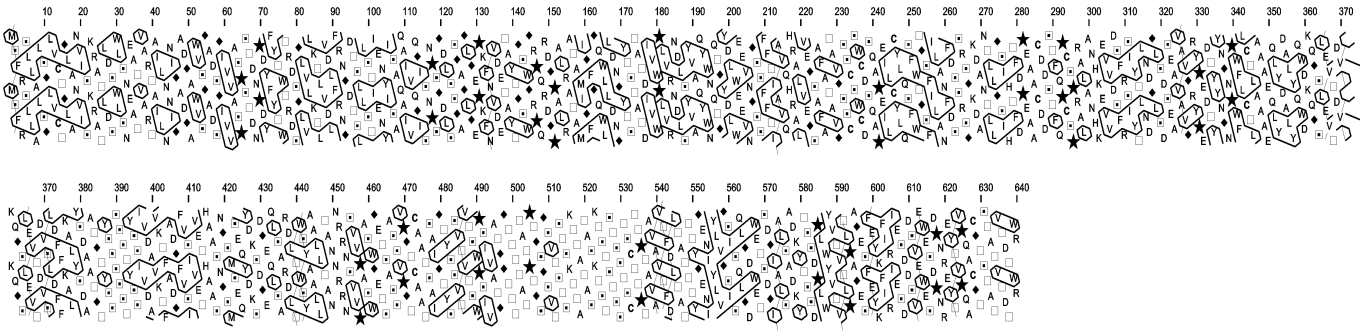
Streptomyces griseus α-amylase



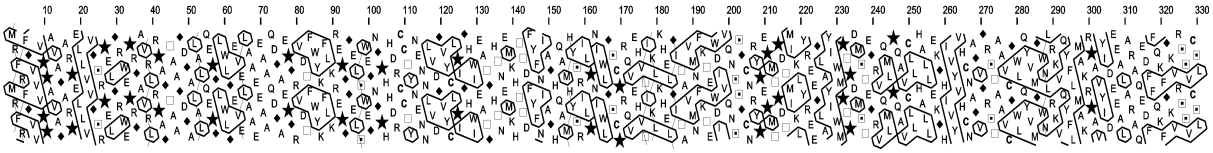
Bacillus cereus β-amylase



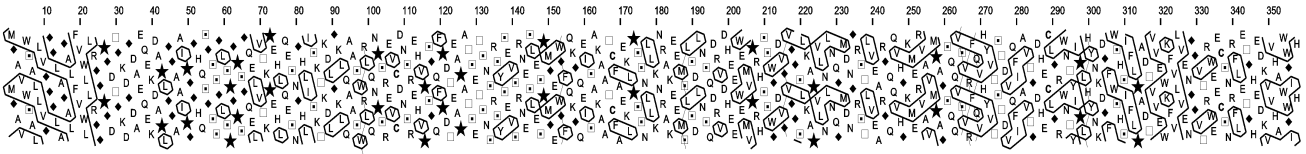
Aspergillus niger glucoamylase



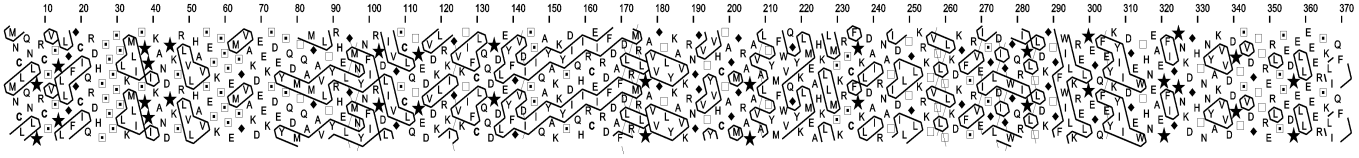
Homo sapiens laforin



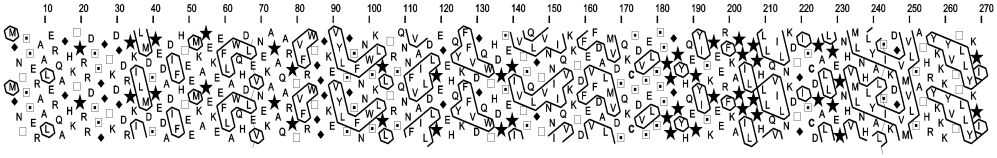
Homo sapiens genethonin-1



Arabidopsis thaliana SEX-4



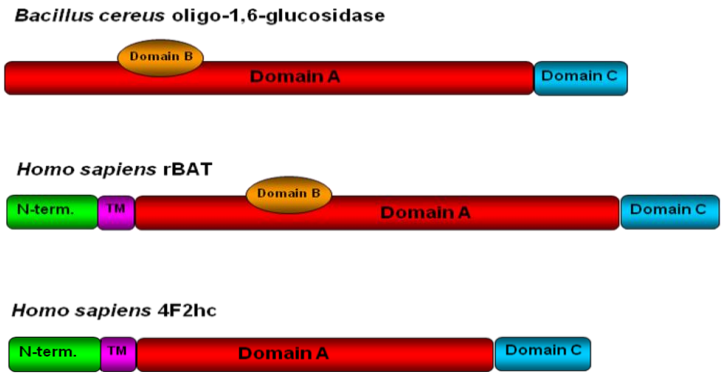
Rattus norvegicus AMPK subunit β1



7.8. Analýza α-glukozidáz s využitím nástroja BLAST

Teória

Enzým oligo-1,6-glukozidáza (OGLU), EC 3.2.1.10, patrí medzi glykozidové hydrolázy. Ako α-glukozidáza katalyzuje hydrolýzu α-1,6-glukozidových väzieb v α-glukánoch. V rámci α-amylázovej enzýmovej rodiny GH13 tvorí tzv. oligo-1,6-glukozidázovú podrodinu spolu s niekoľkými ďalšími príbuznými enzýmovými špecificitami. Tieto enzýmy sú prevažne mikrobiálneho pôvodu. Bolo zistené, že k oligo-1,6-glukozidáze a enzýmom z jej podrodiny existujú u vyšších organizmov (živočíchy až cicavce) sekvenčne podobné proteíny, ktoré tvoria ťažké reťazce heteromérnych transportérov zodpovedných za prenos aminokyselín cez bunkovú membránu, tzv. proteíny rBAT a 4F2hc antigén. Oligo-1,6-glukozidáza je multidoménový enzým, zložený z troch domén (obr. 7.8): (i) N-terminálna katalytická doména A; (ii) doména B včlenená v katalytickej doméne; a (iii) C-terminálna doména C. Transportné proteíny majú podobné usporiadanie domén svojich proteínových molekúl, ale majú navyše N-terminálnu doménu (N-term.) a transmembránový segment (TM). Doména A v proteínoch rBAT a 4F2hc nemusí obsahovať katalytické zvyšky nevyhnutné pre funkciu α-glukozidáz. Proteín 4F2hc navyše neobsahuje ani doménu B.



Obr. 7.8. Schéma oligo-1,6-glukozidázy a proteínov rBAT a 4F2hc.

Zadanie

(1) *In silico* analýza α-glukozidáz z oligo-1,6-glukozidázovej podrodiny pomocou nástroja BLAST. Query: OGLU z *Bacillus cereus* (UniProt: P21332; dĺžka: 558); maximálny počet zachytených sekvencií: 500.

(2) Vytvorte si priečinok „BLAST_OGLU“.

(3) Z databázy UniProt získajte aminokyselinovú sekvenciu OGLU (výsledok uložte).

(4) So sekvenciou danej OGLU vykonajte štandardný proteínový BLAST (Max. target sequences: 500); kompletné výsledky uložte.

(5) V predpripravených výsledkoch získaných pomocou nástroja BLAST (Max. target sequences: 5000) dostupných na web-stránke: http://imb.savba.sk/~janecek/UCM/Tretiaci/BLAST/OGLU/OGLU_BLAST_5000.htm identifikujte ďalších 21 enzýmov/proteínov (č. 2-22; tab. 7.8) a doplňte všetky príslušné údaje (vzorová tabuľka je tiež k dispozícii): http://imb.savba.sk/~janecek/UCM/Tretiaci/BLAST/OGLU/OGLU_Tabulka_vzor.doc

(6) Údaje pre 4F2hc antigén a rBAT proteíny (č. 23, 24 a 25; tab. 7.8) doplňte priamo z databázy GenBank (predpripravený BLAST pre 5000 zachytených sekvencií ich neobsahuje).

(7) Všetkých 25 sekvencií zhromaždite z databáz GenBank a/alebo UniProt do vstupného súboru „OGLU.txt“.

(8) Na EBI serveri zrovnajte sekvencie v programe Clustal-Omega – získajte dva súbory: „OGLU_aln.txt“ (alignment „with character counts“) a „OGLU_fas.txt“ (formát Pearson/FASTA); dodržte vstupné poradie sekvencií (input order).

(9) V zrovnaných sekvenciách, t.j. v súbore „OGLU_aln.txt“ uloženého ako dokument súboru „OGLU_aln.doc“, žltým podsvietením (ak sú prítomné) farebne zvýraznite charakteristické konzervované sekvenčné regióny (udané v sekvencii oligo-1,6-glukozidázy z *Bacillus cereus*), pričom sa zvlášť zamerajte na prítomnosť, resp. neprítomnosť troch katalytických zvyškov (najmä u transportných proteínov; hrubo označené kurzívou): 44_GIDVIWLSP ... 98_DLVVNH ... 167_QPDLN ... 195_GFRM***D***VINF ... 251_MTVG***E***MPG ... 324_YWNNH***D*** ... 360_GTPYIYQGE.

(10) Na EBI serveri v rámci programu Simple-Phylogeny vypočítajte dva evolučné stromy pre študované enzýmy a proteíny – s uvažovaním, resp. s ignorovaním medzier v sekvenciách „OGLU_off.txt“ a „OGLU_on.txt“. Evolučné stromy zobrazte v programe iTOL („circular“), exportujte ako PNG súbory a vložte pod zrovnanie do súboru „OGLU_aln.doc“.

(11) Pre zrovnané sekvencie (alignment) zistite konsenzuálnu dĺžku (CL), vypočítajte sekvenčnú identitu (SI) a podobnosť (SS) a výsledky spolu s diskusiou celej práce slovne zapíšte.

(12) V priečinku „BLAST_OGLU“ budú uložené nasledovné súbory:

(a) celý html súbor OGLU z databázy UniProt („OGLU_seq.htm“);

(b) Vaše výsledky z BLASTu (html súbor: „OGLU_BLAST.htm“; Max. target sequences: 500);

(c) doplnená tab. 7.8 s 25 zdrojmi (DOC súbor: „OGLU_Tabulka.doc“);

(d) vstupný súbor s 25 sekvenciami (textový súbor: „OGLU.txt“);

(e) zrovnania sekvencií (súbory: „OGLU_aln.txt“ a „OGLU_fas.txt“);

(f) zrovnanie vo formáte ALN („OGLU_aln.doc“) so žlto zvýraznenými konzervovanými regiónmi a červenou inverziou zvýraznenými prípadnými katalytickými zvyškami (podľa bodu 9), vloženými obrázkami evolučných stromov (podľa bodu 10) a pod tým hodnoty CL, SI a SS a diskusia evolučných stromov;

(g) vypočítané súbory (nie obrázky) evolučných stromov („OGLU_off.txt“ a „OGLU_on.txt“).

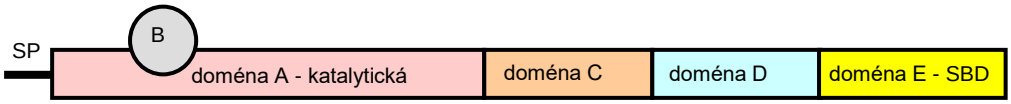
Tabuľka 7.8. Zoznam študovaných sekvencií α-glukozidáz.

Č.	Zdroj enzýmu	Enzým	Skratka	EC	Prístupové číslo	Dĺžka	Skóre
1	*Bacillus cereus*	Oligo-1,6-glukozidáza	Bacce_OGLU	3.2.1.10	sp P21332	558	1152
2		Oligo-1,6-glukozidáza	Bacth_OGLU	3.2.1.10	sp P29094		
3		Oligo-1,6-glukozidáza	Grbha_OGLU	3.2.1.10	gb ENH95997.1		
4		Oligo-1,6-glukozidáza	Anbfl_OGLU	3.2.1.10	gb GAC90225.1		
5		Oligo-1,6-glukozidáza	Bacsu_OGLU	3.2.1.10	gb AAG23399.1		
6		Trehalóza-6-fosfát hydroláza	Anbfl_T6PH	3.2.1.93	gb GAC89934.1		
7		Trehalóza-6-fosfát hydroláza	Gebst_T6PH	3.2.1.93	gb BAE45038.1		
8		Oligo-1,6-glukozidáza	Bacco_OGLU	3.2.1.10	sp Q45101		
9		Oligo-1,6-glukozidáza	Clocl_OGLU	3.2.1.10	gb ENZ09185.1		
10		Alfa-glukozidáza	Labpe_AGLU	3.2.1.20	gb CCC16900.1		
11		Oligo-1,6-glukozidáza	Clobo_OGLU	3.2.1.10	gb ENZ38195.1		
12		Oligo-1,6-glukozidáza	Arbgl_OGLU	3.2.1.10	gb BAC78840.1		
13		Alfa-glukozidáza	Gebst_AGLU	3.2.1.20	gb BAA12704.1		
14		Glykozidáza	Jabli_GLY	3.2.1.-	gb AAZ39207.1		
15		Glykozidáza	Rucgn_GLY	3.2.1.-	gb CCG93502.1		
16		Alfa-glukozidáza	Pecpe_AGLU	3.2.1.20	sp P43473		
17		Oligo-1,6-glukozidáza	Weith_OGLU	3.2.1.10	gb CCC56207.1		
18		Oligo-1,6-glukozidáza	Enbpu_OGLU	3.2.1.10	gb CAZ90594.1		
19		Hypotetický proteín	Cloha_HYPO	----	gb ENY93903.1		
20		Alfa-glukozidáza	Stcau_AGLU	3.2.1.20	gb ENM74527.1		
21		Oligo-1,6-glukozidáza	Bacsp_OGLU	3.2.1.10	sp P29093		
22		Oligo-1,6-glukozidáza	Rhisp_OGLU	3.2.1.10	gb CCF21491.1		
23		4F2hc antigén (isoform 2)	Homsa_4F2	----	gb AAA35489.1		----
24		rBAT proteín	Salsa_rBAT	----	gb ACN60293.1		----
25		rBAT proteín	Homsa_rBAT	----	gb AAA35500.1		----

7.9. Analýza škrob-viažucej domény amyláz s využitím nástroja BLAST

Teória

Enzým cyklodextrínglukanotransferáza (CGTáza), EC 2.4.1.19, patrí medzi transferázy. Katalyzuje hydrolytické štiepenie α-1,4-glukozidových väzieb (napr. v škrobe), pričom súčasne prenáša (transferázová aktivita) molekulu glukózy s následnou tvorbou cyklického sacharidu (cyklodextrínu). Vo všeobecnosti cyklizuje časť α-1,4-D-glukánového reťazca tvorbou α-1,4-D-glukozidovej väzby. CGTáza je multidoménový enzým, ktorý sa najčastejšie skladá z piatich domén (obr. 7.9). Na C-konci svojej proteínovej molekuly obsahuje tzv. škrob-viažucu doménu (starch-binding domain; SBD), ktorá je klasifikovaná ako tzv. modul z rodiny CBM20, t.j. „carbohydrate-binding module family 20“. Táto doména umožňuje enzýmu viazať a degradovať surový (tepelne neupravený) škrob. CBM20 sa môže nachádzať aj v α-amylázach, β-amylázach a glukoamylázach. Hoci tieto jednotlivé amylázy obsahujú sekvenčne nepodobné katalytické domény, na svojich C-koncoch môžu mať odpovedajúcu SBD typu CBM20. Doména SBD typu CBM20 a jej homológ z rodiny CBM48 sa môžu nachádzať napr. v glukánfosfatázach zapojených do regulácie metabolizmu glykogénu u živočíchov (proteín laforín) a škrobu u rastlín (tzv. proteín nadbytku škrobu 4 – SEX-4), prípadne ako súčasť β-podjednotky AMP-aktivovanej proteínovej kinázy (AMPKβ).



Obr. 7.9. Schéma CGTázy (SP, signálny peptid).

Zadanie

(1) *In silico* analýza škrob-viažucich domén z rodiny CBM20 pomocou nástroja BLAST. Query: SBD CGTázy z *Bacillus circulans* (UniProt: P30920; dĺžka: 718; SBD: 613-718); maximálny počet zachytených sekvencií: 250.

(2) Vytvorte si priečinok „BLAST_CBM20“.

(3) Z databázy UniProt získajte aminokyselinovú sekvenciu SBD z danej CGTázy (výsledok uložte).

(4) So sekvenciou danej SBD vykonajte štandardný proteínový BLAST (Max. target sequences: 250); kompletné výsledky uložte.

(5) Na výber databázy si (buď sami alebo po dohode s vyučujúcim) zvoľte niektorú vhodnú kombináciu pre „Organism“ a „Exclude“ z nasledovných možností pri vylúčení taxónu „Paenibacillus (taxid:44249)“: „Bacteria (taxid:2)“; „Archaea (taxid:2157)“; „Plants (taxid:3193)“; „Viridiplantae (taxid:33090)“; „Eucarya (taxid:2759)“;

„Fungi (taxid:4751)“; „Fungi/Metazoa group (taxid:33154)“; „Metazoa (taxid:33208)“ a Mammals (taxid:40674).

(6) Vo výsledkoch získaných BLASTom identifikujte (na základe tab. 7.9) ďalších 29 enzýmov/proteínov spomedzi CGTáz, prípadne α-amyláz, β-amyláz a glukoamyláz, ako aj ďalších príbuzných hypotetických enzýmov a proteínov (t.j. dohromady so vstupnou SBD 30), ktoré budú obsahovať motív SBD (~ 100 aminokyselinových zvyškov) podľa nasledovného vzoru: ~30 aa ... **lg-W** ... ~15 aa ... **p-W** ... ~15 aa ... **fKf** ... ~10 aa ... **W----n** ... ~15 aa (C-koniec domény). Požadované údaje – extrahované z výsledkov BLASTu – vhodne uložte do tab. 7.9.

(7) Všetkých 30 sekvencií SBD – vystrihnutých z pôvodných sekvencií celých enzýmov a proteínov – zhromaždite z databáz GenBank a/alebo UniProt do vstupného súboru „SBD.txt“.

(8) Na EBI serveri zrovnajte sekvencie SBD v programe Clustal-Omega – získajte dva súbory: „SBD_aln.txt“ (alignment „with character counts“) a „SBD_fas.txt“ (formát Pearson/FASTA); dodržte vstupné poradie sekvencií (input order).

(9) V zrovnaných SBD sekvenciách, t.j. v súbore „SBD_aln.txt“ uloženého ako dokument súboru „SBD_aln.doc“, žltým podsvietením farebne zvýraznite charakteristické konzervované zvyšky (ak sú prítomné): ~30 aa ... lg-W ... ~15 aa ... p-W ... ~15 aa ... fKf ... ~10 aa ... W----n ... ~15 aa (C-koniec zrovnania).

(10) Na EBI serveri v rámci programu Simple-Phylogeny vypočítajte dva evolučné stromy pre SBD – s uvažovaním, resp. s ignorovaním medzier v sekvenciách „SBD_off.txt“ a „SBD_on.txt“. Evolučné stromy zobrazte v programe iTOL („circular“), exportujte ako PNG súbory a vložte pod zrovnanie do súboru „SBD_aln.doc“.

(11) Pre zrovnané SBD sekvencie (alignment) zistite konsenzuálnu dĺžku (CL), vypočítajte sekvenčnú identitu (SI) a podobnosť (SS) a výsledky spolu s diskusiou celej práce slovne zapíšte.

(12) V priečinku „BLAST_CBM20“ budú uložené nasledovné súbory:

(a) celý html súbor CGTázy z databázy UniProt („CGTase.htm);

(b) Vaše výsledky z BLASTu (html súbor: „SBD_BLAST.htm“; Max. target sequences: 250);

(c) vyplnená tab. 7.9 s 30 zdrojmi SBD identifikovanými v BLASte (DOC súbor: „SBD_Tabulka.doc);

(d) vstupný súbor s 30 SBD sekvenciami (textový súbor: „SBD.txt“);

(e) zrovnania SBD sekvencií (súbory: „SBD_aln.txt“ a „SBD_fas.txt“);

(f) zrovnanie vo formáte ALN („SBD_aln.doc“) s farebne zvýraznenými konzervovanými zvyškami (podľa bodu 9), vloženými obrázkami evolučných stromov (podľa bodu 10) a pod tým hodnoty CL, SI a SS a diskusia evolučných stromov;

(g) vypočítané súbory (nie obrázky) evolučných stromov („SBD_off.txt“ a „SBD_on.txt“).

Tabuľka 7.9. Zoznam sekvencií škrob-viažucich domén z rodiny CBM20 (prípadne CBM48).

Č.	Skratka	Enzým/proteín	Zdroj	UniProt/GenBank	E-hodnota	Skóre	Dĺžka	SBD
1	Bacci_CGT	CGTáza	*Bacillus circulans*	P30920			718	613-718
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								

7.10. Predikcia a porovnávanie terciárnej štruktúry proteínov

Teória

Proteín:
- α-amyláza z *Haloarcula hispanica* z rodiny glykozidových hydroláz GH13;
- UniProt Acc. No.: Q4A3E0;
- dĺžka: 433 aminokyselinových zvyškov;
- štruktúra je zložená z troch domén: (i) doména A – katalytická TIM-barelová doména: časť A1 – Met1-Asn136 a časť A2 – Ala193-Ser355; (ii) doména B – v pozícii medzi vláknom β3 a helixom α3: His137-Ser192; a (iii) doména C – antiparalelná β-sendvičová C-terminálna doména: Gly356-Glu433;
- potenciálne katalytické zvyšky: Asp217 (katalytický nukleofil), Glu245 (donor protónov) a Asp308 (stabilizátor prechodového stavu).

Zadanie:

(1) Vytvorte si priečinok „Haloarcula“.

(2) Pre sekvenciu α-amylázy z *Haloarcula hispanica* vykonajte predikciu terciárnej štruktúry metódou rozpoznania potenciálnej terciárnej štruktúry (tzv. „fold recognition“) na serveri Phyre2: www.sbg.bio.ic.ac.uk/phyre2/. Všetkých 20 získaných modelov uložte ako PDB súbory „AAMY_Halhi_xxxx_1.pdb“, „AAMY_Halhi_xxxx_2.pdb“, …, „AAMY_Halhi_xxxx_20.pdb“; kde „xxxx“ je PDB kód terciárnej štruktúry templátu a čísla 1-20 sú poradové čísla jednotlivých modelov.

(3) Vyberte si 7 modelov (nie prvých 7 modelov v poradí) terciárnej štruktúry α-amylázy z *Haloarcula hispanica* a zobrazte ich v programe WebLabViewerLite pri vypnutí atómov ako „solid ribbon“, pričom jednotlivé domény farbite nasledovne: doména A – modrá, doména B – červená a doména C – zelená. Potom zobrazte katalytické aminokyselinové zvyšky ako „stick“ a farbite ako „element“. Všetkých 7 zobrazených modelov sa pokúste kvôli vizuálnemu porovnaniu naorientovať podobne a uložte ako „*.msv“ aj „*.gif“, resp. „*.jpg“ súbory.

(4) Máte k dispozícii, resp. získajte:

   (a) koordináty 3 modelov terciárnej štruktúry α-amylázy z *Haloarcula hispanica*, získané metódou homologického modelovania (Phyre2 server); tieto súbory označte: „AAMY_Phyre_1HVX.pdb“, „AAMY_Phyre_3BLP.pdb“ a „AAMY_Phyre_1MWO.pdb“;

   (b) koordináty 3 experimentálne určených terciárnych štruktúr reálnych α-amyláz z *Bacillus stearothermophilus* (PDB: 1HVX), z *Homo sapiens* (PDB: 3BLP) a *Pyrococus woesei* (PDB: 1MWO); tieto súbory označte: „AAMY_1HVX.pdb“, „AAMY_3BLP.pdb“ a „AAMY_1MWO.pdb“.

(5) Vytvorte si 9 priečinkov a v nich 9 ZIP súborov obsahujúcich všetky príslušné dvojice štruktúr modelu α-amylázy z *H. hispanica* a reálnej štruktúry podľa nasledovného členenia:

- 1: AAMY_Phyre_1HVX.pdb a AAMY_1HVX.pdb;
- 2: AAMY_Phyre_1HVX.pdb a AAMY_3BLP.pdb;
- 3: AAMY_Phyre_1HVX.pdb a AAMY_1MWO.pdb;
- 4: AAMY_Phyre_3BLP.pdb a AAMY_1HVX.pdb;
- 5: AAMY_Phyre_3BLP.pdb a AAMY_3BLP.pdb;
- 6: AAMY_Phyre_3BLP.pdb a AAMY_1MWO.pdb;
- 7: AAMY_Phyre_1MWO.pdb a AAMY_1HVX.pdb
- 8: AAMY_Phyre_1MWO.pdb a AAMY_3BLP.pdb
- 9: AAMY_Phyre_1MWO.pdb a AAMY_1MWO.pdb.

(6) Na serveri MultiProt: http://bioinfo3d.cs.tau.ac.il/MultiProt/ podľa požiadaviek („upload a ZIP file of PDB structures“) spravte postupne vzájomné preloženia všetkých 9 dvojíc štruktúr a poskytnuté údaje zapíšte do tab. 7.10 – hodnotu RMSD a počet korešpondujúcich Cα atómov (v zátvorke):

Tabuľka 7.10. Superpozície modelov α-amylázy z *H. hispanica* s templátmi.

	AAMY_Phyre_1HVX.pdb	AAMY_Phyre_3BLP.pdb	AAMY_Phyre_1MWO.pdb
AAMY_1HVX.pdb			
AAMY_3BLP.pdb			
AAMY_1MWO.pdb			

(7) Všetkých 9 preložených štruktúr (PDB alignment) spolu s príslušným zrovnaním (Alignment size) a výsledkami uložte s názvom ZIP súboru (napr. „1_overlap.pdb“, „1_alignment.htm“ a „1_vysledky.htm“ – ako kompletnú web-stránku; atď.). Zrovnania vizuálne prezrite s ohľadom na funkčne dôležité zvyšky (udané nižšie) a preklady zobrazte v programe WebLabViewerLite pri vypnutí atómov ako „solid ribbon“ a farbite vždy jednu štruktúru tyrkysovou a druhú oranžovou farbou. Potom zobrazte katalytické a funkčne dôležité aminokyselinové zvyšky – ako „stick“: (i) α-amyláza z *Haloarcula hispanica* – His137, Arg215, Asp217, Glu245, His307 a Asp308; (ii) α-amyláza z *Bacillus stearothermophilus* – His106, Arg232, Asp234, Glu264, His330 a Asp331; (iii) α-amyláza z *Homo sapiens* – His101, Arg195, Asp197, Glu233, His299 a Asp300; a (iv) α-amyláza z *Pyrococus woesei* – His111, Arg196, Asp198, Glu222, His288 a Asp289; a farbite vždy modrou v tyrkysovej štruktúre a fialovou v oranžovej štruktúre. Všetky zobrazenia uložte ako príslušné „*.msv“ a „*.gif“, resp. „*.jpg“ súbory.

7.11. Projekt kompletnej bioinformatickej analýzy proteínu/enzýmu

„Acc_No“ predstavuje prístupové číslo (accession No.) ľubovoľného proteínu/enzýmu z databázy UniProt.

(1) Uložte kompletný súbor pre Váš proteín/enzým z databázy UniProt (http://www.uniprot.org/) ako HTML súbor: „Acc_No.html“ (kompletná web-stránka).

(2) Identifikujte tento proteín/enzým v databáze GenBank (http://www.ncbi.nlm.nih.gov/genbank/) a uložte kompletný súbor pre jeho gén („genomic DNA“), ako aj preklad („translation“; „protein_id“) ako HTML súbory: „GenBank_Acc_No.html“ a „GenPept_Acc_No.html“ (kompletné web-stránky). Nájdite v literatúre publikáciu, v ktorej bola popísaná sekvencia Vášho proteínu/enzýmu a súbor uložte ako PDF súbor: „Acc_No.pdf“.

(3) So sekvenciou Vášho proteínu/enzýmu vykonajte štandardný proteínový BLAST (Max. target sequences: 500; http://blast.ncbi.nlm.nih.gov/Blast.cgi) a výsledky uložte ako HTML súbor: „Acc_No_BLAST_500.html“ (kompletná web-stránka). Vo výsledkoch BLASTu identifikujte ďalších 39 príbuzných proteínov/enzýmov do tabuľky (názov: „Tabuľka študovaných proteínov/enzýmov“; uložená ako DOC súbor: „Acc_No_Table.doc“), v ktorej budú nasledovné stĺpce: (i) „Číslo“; (ii) „Zdroj“ (proteínu/enzýmu); (iii) „Skratka“; (iv) „UniProt“ (accession No.); (v) „GenBank“ (accession No.); (vi) „Dĺžka“ (proteínu/enzýmu); a (vii) „Skóre“ (z BLASTu). To znamená, že Váš proteín/enzým bude č. 1 v tabuľke a pod ním bude ďalších 39 proteínov/enzýmov usporiadaných podľa klesajúceho skóre z BLASTu. Pokiaľ sa to bude dať, pokúsite sa identifikovať tých 39 príbuzných proteínov/enzýmov z čo najširšieho spektra organizmov (*Archaea*, *Bacteria*, *Eucarya*).

(4) Sekvencie všetkých 40 proteínov/enzýmov porovnajte, to znamená, že: (i) pripravíte vstupný súbor („Acc_No_input.txt“); (ii) na EBI serveri (http://www.ebi.ac.uk/) sekvencie zrovnáte v programe Clustal-Omega – získate dva súbory: „Acc_No_input_aln.txt“ a „Acc_No_input_fas.txt“ – zrovnania vo formáte „alignment“ (Clustal with character counts) a „Pearson/FASTA“; (iii) na EBI serveri v rámci programu Simple-Phylogeny vypočítate dva evolučné stromy pre študované proteíny/enzýmy – s uvažovaním medzier v sekvenciách („Acc_No_input_off.txt“) a s ich ignorovaním („Acc_No_input_on.txt“); (iv) zrovnanie vo formáte „alignment“ uložte ako DOC („Acc_No_input_aln.doc“) a v ňom zvýraznite žltým podfarbením funkčne dôležité zvyšky (napr. katalytické zvyšky, prípadne zvyšky aktívneho miesta) identifikované buď z databázy UniProt alebo z literatúry, prípadne z výsledkov BLASTu, ďalej pod zrovnaním budú zapísané hodnoty konsenzuálnej dĺžky (CL), sekvenčnej identity (SI) a sekvenčnej podobnosti (SS), vložené obrázky evolučných stromov (napr. ako PNG)

získané v programe iTOL (interactive Tree of Life; https://itol.embl.de/) a ich slovná diskusia.

(5) Pre sekvenciu Vášho proteínu/enzýmu vykonajte predikciu sekundárnej štruktúry metódou GOR-IV (na serveri http://npsa-prabi.ibcp.fr/). Kompletné výsledky z predikcie serveru uložte ako HTML súbor: „Acc_No_2D_GOR4_results.html“ (kompletná web-stránka). Pravidelné elementy získanej predpovedanej sekundárnej štruktúry (t.j. iba H a E, t.j. α-helix a β-list) zapíšte pod aminokyselinovú sekvenciu a uložte v riadkoch po 60 pozícií do textového súboru: „Acc_No_2D_GOR4.txt“ (1. riadok – poradové číslo v sekvencii: 10, 20, 30, atď; 2. riadok – aminokyselinová sekvencia; a 3. riadok – predikcia H, resp. E – pod príslušným zvyškom).

(6) Pre sekvenciu Vášho proteínu/enzýmu vykonajte predikciu terciárnej štruktúry metódou rozpoznania terciárnej štruktúry (tzv. „fold recognition“) na serveri Phyre2: http://www.sbg.bio.ic.ac.uk/phyre2/. Po prezretí výsledkov vyberte jeden najlepší model a uložte ho ako PDB súbor: „Acc_No_3D_Phyre2_template.pdb“ (namiesto „template“ bude vždy v názve súboru PDB-kód templátovej terciárnej štruktúry, podľa ktorej bol model Vášho proteínu/enzýmu získaný). Váš výber modelu z výsledkov predikcie zdôvodnite (podľa „alignment coverage“, „confidence“, „% i.d.“ a „template information“) v textovom súbore: „Acc_No_3D_Phyre2_template.txt“, kde bude aj informácia, koľko templátov bolo identifikovaných a koľko modelov bolo skutočne pripravených, ako aj www adresa, kde sa na Phyre2 serveri dočasne nachádzajú kompletné výsledky predikcie Vášho proteínu/enzýmu. Z databázy PDB (http://www.rcsb.org/) získajte koordináty terciárnej štruktúry templátu a uložte ako PDB súbor: „Acc_No_PDB_template.pdb“. Obe štruktúry, t.j. aj model Vášho proteínu/enzýmu, aj templát zobrazte v programe WebLabViewerLite ako pohľad na kompletnú terciárnu štruktúru pri zobrazení „solid ribbon“ (pri súčasnom vypnutí atómov) a funkčne dôležité aminokyselinové zvyšky (napr. katalytické, resp. zvyšky aktívneho miesta) zobrazte ako „stick“ a farbite ako „element“ – súbory „*.msv“: (i) Váš model – „Acc_No_3D_Phyre2_template.msv“; a (ii) templát – „Acc_No_PDB_template.msv“.

(7) Pripravte si ZIP súbor „Acc_No_MultiProt.zip“ so štruktúrou Vášho modelu a jeho templátu a na serveri MultiProt (http://bioinfo3d.cs.tau.ac.il/MultiProt/) podľa požiadaviek servera spravte vzájomnú superpozíciu oboch štruktúr. Poskytnuté údaje – hodnotu RMSD a počet korešpondujúcich Cα atómov (v zátvorke) – zapíšte do textového súboru „Acc_No_MultiProt_results.txt“. Získané výsledky uložte nasledovne: (i) stránku s výsledkami ako „Acc_No_MultiProt_results.html“ (kompletná web-stránka); (ii) zrovnanie sekvencií založené na štruktúre – „alignment size“ ako „Acc_No_MultiProt_alignment.html“ (kompletná web-stránka); a (iii) preložené štruktúry – „PDB alignment“ ako PDB súbor „Acc_No_Multiprot_overlap.pdb“. Obe preložené štruktúry zobrazte v programe WebLabViewerLite ako: (i) celkový pohľad na kompletnú

terciárnu štruktúru pri zobrazení „solid ribbon“ (pri súčasnom vypnutí atómov) s funkčne dôležitými aminokyselinovými zvyškami (napr. katalytickými, resp. zvyškami aktívneho miesta) ako „stick“ (súbor „Acc_No_Multiprot_overlap.msv“); a (ii) detailný pohľad (priblíženie) na oblasť s funkčne dôležitými zvyškami v zobrazení proteínu/enzýmu ako „line ribbon“ a selektovanými zvyškami ako „stick“ (súbor „Acc_No_Multiprot_overlap_detail.msv“). Model Vášho proteínu/enzýmu farbite tyrkysovou farbou a štruktúru templátu farbite oranžovou farbou, pričom funkčne dôležité zvyšky farbite modrou farbou v modeli a červenou farbou v templáte.

(8) Výsledky tohto projektu budú pripravené ako ZIP, ktorý bude obsahovať tieto súbory:

(a) Acc_No.html
(b) GenBank_Acc_No.html
(c) GenPept_Acc_No.html
(d) Acc_No.pdf
(e) Acc_No_BLAST_500.html
(f) Acc_No_Table.doc
(g) Acc_No_input.txt
(h) Acc_No_input_aln.txt
(i) Acc_No_input_fas.txt
(j) Acc_No_input_off.txt
(k) Acc_No_input_on.txt
(l) Acc_No_input_aln.doc
(m) Acc_No_2D_GOR4_results.html
(n) Acc_No_2D_GOR4.txt
(o) Acc_No_3D_Phyre2_template.pdb
(p) Acc_No_3D_Phyre2_template.txt
(q) Acc_No_PDB_template.pdb
(r) Acc_No_3D_Phyre2_template.msv
(s) Acc_No_PDB_template.msv
(t) Acc_No_MultiProt.zip
(u) Acc_No_MultiProt_results.txt
(v) Acc_No_MultiProt_results.html
(w) Acc_No_MultiProt_alignment.html
(x) Acc_No_Multiprot_overlap.pdb
(y) Acc_No_Multiprot_overlap.msv
(z) Acc_No_Multiprot_overlap_detail.msv

Použitá a odporúčaná literatúra

Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul S.F. & Koonin E.V. 1998. Iterated profile searches with PSI-BLAST – a tool for discovery in protein databases. *Trends Biochem. Sci.* **23:** 444–447. https://doi.org/10.1016/s0968-0004(98)01298-5

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W. & Lipman D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amid C., Alako B.T.F., Balavenkataraman Kadhirvelu V., Burdett T., Burgin J., Fan J., Harrison P.W., Holt S., Hussein A., Ivanov E., Jayathilaka S., Kay S., Keane T., Leinonen R., Liu X., Martinez-Villacorta J., Milano A., Pakseresht A., Rahman N., Rajan J., Reddy K., Richards E., Smirnov D., Sokolov A., Vijayaraja S. & Cochrane G. 2020. The European Nucleotide Archive in 2019. *Nucleic Acids Res.* **48 (Database Issue 1):** D70–D76. https://doi.org/10.1093/nar/gkz1063

Armstrong D.R., Berrisford J.M., Conroy M.J., Gutmanas A., Anyango S., Choudhary P., Clark A.R., Dana J.M., Deshpande M., Dunlop R., Gane P., Gáborová R., Gupta D., Haslam P., Koča J., Mak L., Mir S., Mukhopadhyay A., Nadzirin N., Nair S., Paysan-Lafosse T., Pravda L., Sehnal D., Salih O., Smart O., Tolchard J., Varadi M., Svobodova-Vařeková R., Zaki H., Kleywegt G.J. & Velankar S. 2020. PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.* **48 (Database Issue 1):** D335–D343. https://doi.org/10.1093/nar/gkz990

Blattner F.R., Plunkett G. 3rd, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B. & Shao Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:** 1453–1462. https://doi.org/10.1126/science.277.5331.1453

Boratyn G.M., Schäffer A.A., Agarwala R., Altschul S.F., Lipman D.J. & Madden T.L. 2012. Domain enhanced lookup time accelerated BLAST. *Biol. Direct* **7:** 12. https://doi.org/10.1186/1745-6150-7-12

Bult C.J., White O., Olsen G.J., Zhou L., Fleischmann R.D., Sutton G.G., Blake J.A., FitzGerald L.M., Clayton R.A., Gocayne J.D., Kerlavage A.R., Dougherty B.A., Tomb J.F., Adams M.D., Reich C.I., Overbeek R., Kirkness E.F., Weinstock K.G., Merrick J.M., Glodek A., Scott J.L., Geoghagen N.S. & Venter J.C. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* **273:** 1058–1073. https://doi.org/10.1126/science.273.5278.1058

Callebaut I., Labesse G., Durand P., Poupon A., Canard L., Chomilier J., Henrissat B. & Mornon J.P. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cell. Mol Life Sci.* 53: 621–645. https://doi.org/10.1007/s000180050082

Cole S.T., Brosch R., Parkhill J., Garnier T., Churcher C., Harris D., Gordon S.V., Eiglmeier K., Gas S., Barry C.E. 3rd, Tekaia F., Badcock K., Basham D., Brown D., Chillingworth T., Connor R., Davies R., Devlin K., Feltwell T., Gentles S., Hamlin N., Holroyd S., Hornsby T., Jagels K., Krogh A., McLean J., Moule S., Murphy L., Oliver K., Osborne J., Quail M.A., Rajandream M.A., Rogers J., Rutter S., Seeger K., Skelton J., Squares R., Squares S., Sulston J.E., Taylor K., Whitehead S. & Barrell B.G. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393:** 537–544. https://doi.org/10.1038/31159

Cook C.E., Stroe O., Cochrane G., Birney E. & Apweiler R. 2020. The European Bioinformatics Institute in 2020: building a global infrastructure of interconnected data resources for the life sciences. *Nucleic Acids Res.* **48 (Database Issue 1):** D17–D23. https://doi.org/10.1093/nar/gkz1033

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M., McKenney K., Sutton G.G., FitzHugh W., Fields C.A., Gocayne J.D., Scott J.D., Shirley R., Liu L.I., Glodek A., Kelley J.M., Weidman J.F., Phillips C.A., Spriggs T., Hedblom E., Cotton M.D., Utterback T., Hanna M.C., Nguyen D.T., Saudek D.M., Brandon R.C., Fine L.D., Fritchman J.L., Fuhrmann J.L., Geoghagen N.S., Gnehm C.L., McDonald L.A., Small K.V., Fraser C.M., Smith H.O. & Venter J.C. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496–512. https://doi.org/10.1126/science.7542800

Fraser C.M., Casjens S., Huang W.M., Sutton G.G., Clayton R., Lathigra R., White O., Ketchum K.A., Dodson R., Hickey E.K., Gwinn M., Dougherty B., Tomb J.F., Fleischmann R.D., Richardson D., Peterson J., Kerlavage A.R., Quackenbush J., Salzberg S., Hanson M., van Vugt R., Palmer N., Adams M.D., Gocayne J., Weidman J., Utterback T., Watthey L., McDonald L., Artiach P., Bowman C., Garland S., Fuji C., Cotton M.D., Horst K., Roberts K., Hatch B., Smith H.O. & Venter J.C. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390:** 580–586. https://doi.org/10.1038/37551

Fraser C.M., Gocayne J.D., White O., Adams M.D., Clayton R.A., Fleischmann R.D., Bult C.J., Kerlavage A.R., Sutton G., Kelley J.M., Fritchman R.D., Weidman J.F., Small K.V., Sandusky M., Fuhrmann J., Nguyen D., Utterback T.R., Saudek D.M., Phillips C.A., Merrick J.M., Tomb J.F., Dougherty B.A., Bott K.F., Hu P.C., Lucier T.S., Peterson S.N., Smith H.O., Hutchison C.A. 3rd & Venter J.C. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270:** 397–403. https://doi.org/10.1126/science.270.5235.397

Fraser C.M., Norris S.J., Weinstock G.M., White O., Sutton G.G., Dodson R., Gwinn M., Hickey E.K., Clayton R., Ketchum K.A., Sodergren E., Hardham J.M., McLeod M.P., Salzberg S., Peterson J., Khalak H., Richardson D., Howell J.K., Chidambaram M., Utterback T., McDonald L., Artiach P., Bowman C., Cotton M.D., Fujii C., Garland S., Hatch B., Horst K., Roberts K., Sandusky M., Weidman J., Smith H.O. & Venter J.C. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281:** 375–388. https://doi.org/10.1126/science.281.5375.375

Gaboriaud C, Bissery V, Benchetrit T, Mornon JP. 1987. Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett.* **224:** 149–155. https://doi.org/10.1016/0014-5793(87)80439-8

Heidelberg J.F., Eisen J.A., Nelson W.C., Clayton R.A., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Umayam L., Gill S.R., Nelson K.E., Read T.D., Tettelin H., Richardson D., Ermolaeva M.D., Vamathevan J., Bass S., Qin H., Dragoi I., Sellers P., McDonald L., Utterback T., Fleishmann R.D., Nierman W.C., White O., Salzberg S.L., Smith H.O., Colwell R.R., Mekalanos J.J., Venter J.C. & Fraser C.M. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406:** 477–483. https://doi.org/10.1038/35020000

Heidelberg J.F., Paulsen I.T., Nelson K.E., Gaidos E.J., Nelson W.C., Read T.D., Eisen J.A., Seshadri R., Ward N., Methe B., Clayton R.A., Meyer T., Tsapin A., Scott J., Beanan M., Brinkac L., Daugherty S., DeBoy R.T., Dodson R.J., Durkin A.S., Haft D.H., Kolonay J.F., Madupu R., Peterson J.D., Umayam L.A., White O., Wolf A.M., Vamathevan J., Weidman J., Impraim M., Lee K., Berry K., Lee C., Mueller J., Khouri H., Gill J., Utterback T.R., McDonald L.A., Feldblyum T.V., Smith H.O., Venter J.C., Nealson K.H. & Fraser C.M. 2002. Genome sequence of the dissimilatory metal ion-reducing bacterium *Shewanella oneidensis*. *Nat. Biotechnol.* **20:** 1118–1123. https://doi.org/10.1038/nbt749

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. https://doi.org/10.1038/35057062

Istrail S., Sutton G.G., Florea L., Halpern A.L., Mobarry C.M., Lippert R., Walenz B., Shatkay H., Dew I., Miller J.R., Flanigan M.J., Edwards N.J., Bolanos R., Fasulo D., Halldorsson B.V., Hannenhalli S., Turner R., Yooseph S., Lu F., Nusskern D.R., Shue B.C., Zheng X.H., Zhong F., Delcher A.L., Huson D.H., Kravitz S.A., Mouchard L., Reinert K., Remington K.A., Clark A.G., Waterman M.S., Eichler E.E., Adams M.D., Hunkapiller M.W., Myers E.W. & Venter J.C. 2004. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl. Acad. Sci. USA* **101:** 1916–1921. https://doi.org/10.1073/pnas.0307971100

Janecek S. 2002. A motif of a microbial starch-binding domain found in human genethonin. *Bioinformatics* **18:** 1534–1537. https://doi.org/10.1093/bioinformatics/18.11.1534

Janecek S. 2014. *Proteínový dizajn.* Učebný text. Univerzita sv. Cyrila a Metoda v Trnave, Trnava. ISBN 978-80-8105-594-2. http://fpv.ucm.sk/images/ucebne_texty/Proteinovy_dizajn.pdf

Janecek S., Marecek F., MacGregor E.A. & Svensson B. 2019. Starch-binding domains as CBM families – history, occurrence, structure, function and evolution. *Biotechnol. Adv.* **37:** 107451. https://doi.org/10.1016/j.biotechadv.2019.107451

Janecek S., Svensson B. & Henrissat B. 1997. Domain evolution in the α-amylase family. *J. Mol. Evol.* **45:** 322–331. https://doi.org/10.1007/pl00006236

Janecek S., Svensson B. & MacGregor E.A. 2003. Relation between domain evolution, specificity, and taxonomy of the α-amylase family members containing a C-terminal starch-binding domain. *Eur. J. Biochem.* **270:** 635–645. https://doi.org/10.1046/j.1432-1033.2003.03404.x

Janecek S., Svensson B. & MacGregor E.A. 2014. α-Amylase – an enzyme specificity found in various families of glycoside hydrolases. *Cell. Mol. Life Sci.* **71:** 1149–1170. https://doi.org/10.1007/s00018-013-1388-z

Janecek S. & Zamocka B. 2020. A new GH13 subfamily represented by the α-amylase from the halophilic archaeon *Haloarcula hispanica*. *Extremophiles* **24:** 207–217. https://doi.org/10.1007/s00792-019-01147-y

Kelley L.A., Mezulis S., Yates C.M., Wass M.N. & Sternberg M.J. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10:** 845–858. https://doi.org/10.1038/nprot.2015.053

Kelley L.A. & Sternberg M.J. 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4:** 363–371. https://doi.org/10.1038/nprot.2009.2

Klein C. & Schulz G.E. 1991. Structure of cyclodextrin glycosyltransferase refined at 2.0 Å resolution. *J. Mol. Biol.* **217:** 737–750. https://doi.org/10.1016/0022-2836(91)90530-j

Klenk H.P., Clayton R.A., Tomb J.F., White O., Nelson K.E., Ketchum K.A., Dodson R.J., Gwinn M., Hickey E.K., Peterson J.D., Richardson D.L., Kerlavage A.R., Graham D.E., Kyrpides N.C., Fleischmann R.D., Quackenbush J., Lee N.H., Sutton G.G., Gill S., Kirkness E.F., Dougherty B.A., McKenney K., Adams M.D., Loftus B., Peterson S., Reich C.I., McNeil L.K., Badger J.H., Glodek A., Zhou L., Overbeek R., Gocayne J.D., Weidman J.F., McDonald L., Utterback T., Cotton M.D., Spriggs T., Artiach P., Kaine B.P., Sykes S.M., Sadow P.W., D'Andrea K.P., Bowman C., Fujii C., Garland S.A., Mason T.M., Olsen G.J., Fraser C.M., Smith H.O., Woese C.R. & Venter J.C. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390:** 364–370. https://doi.org/10.1038/37052

Korenberg J.R., Chen X.N., Adams M.D. & Venter J.C. 1995. Toward a cDNA map of the human genome. *Genomics* **29:** 364–370. https://doi.org/10.1006/geno.1995.9993

Kunst F., Ogasawara N., Moszer I., Albertini A.M., Alloni G., Azevedo V., Bertero M.G., Bessieres P., Bolotin A., Borchert S., Borriss R., Boursier L., Brans A., Braun M., Brignell S.C., Bron S., Brouillet S., Bruschi C.V., Caldwell B., Capuano V., Carter N.M., Choi S.K., Cordani J.J., Connerton I.F., Cummings N.J., Daniel R.A., Denziot F., Devine K.M., Düsterhöft A., Ehrlich S.D., Emmerson P.T., Entian K.D., Errington J., Fabret C., Ferrari E., Foulger D., Fritz C., Fujita M., Fujita Y., Fuma S., Galizzi A., Galleron N., Ghim S.Y., Glaser P., Goffeau A., Golightly E.J., Grandi G., Guiseppi G., Guy B.J., Haga K., Haiech J., Harwood C.R., Henaut A., Hilbert H., Holsappel S., Hosono S., Hullo M.F., Itaya M., Jones L., Joris B., Karamata D., Kasahara Y., Klaerr-Blanchard M., Klein C., Kobayashi Y., Koetter P., Koningstein G., Krogh S., Kumano M., Kurita K., Lapidus A., Lardinois S., Lauber J., Lazarevic V., Lee S.M., Levine A., Liu H., Masuda S., Mauël C., Medigue C., Medina N., Mellado R.P., Mizuno M., Moestl D., Nakai S., Noback M., Noone D., O'Reilly M., Ogawa K., Ogiwara A., Oudega B., Park S.H., Parro V., Pohl T.M., Portelle D., Porwollik S., Prescott A.M., Presecan E., Pujic P., Purnelle B., Rapoport G., Rey M., Reynolds S., Rieger M., Rivolta C., Rocha E., Roche B., Rose M., Sadaie Y., Sato T., Scanlan E., Schleich S., Schroeter R., Scoffone F., Sekiguchi J., Sekowska A., Seror S.J., Serror P., Shin B.S., Soldo B., Sorokin A., Tacconi E., Takagi T., Takahashi H., Takemaru K., Takeuchi M., Tamakoshi A., Tanaka T., Terpstra P., Togoni A., Tosato V., Uchiyama S., Vandebol M., Vannier F., Vassarotti A., Viari A., Wambutt R., Wedler H., Weitzenegger T., Winters P., Wipat A., Yamamoto H., Yamane K., Yasumoto K., Yata K., Yoshida K., Yoshikawa H.F., Zumstein E., Yoshikawa H. & Danchin A. 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390:** 249–256. https://doi.org/10.1038/36786

Lawson C.L., van Montfort R., Strokopytov B., Rozeboom H.J., Kalk K.H., de Vries G.E., Penninga D., Dijkhuizen L. & Dijkstra B.W. 1994. Nucleotide sequence and X-ray structure of cyclodextrin glycosyltransferase from *Bacillus circulans* strain 251 in a maltose-dependent crystal form. *J. Mol. Biol.* **236:** 590–600. https://doi.org/10.1006/jmbi.1994.1168

Lemesle-Varloot L., Henrissat B., Gaboriaud C., Bissery V., Morgat A. & Mornon J.P. 1990. Hydrophobic cluster analysis: procedures to derive structural and functional information from 2-D-representation of protein sequences. *Biochimie* **72:** 555–774. https://doi.org/10.1016/0300-9084(90)90120-6

Lesk A.M. 2001. *Introduction to Protein Architecture.* Oxford University Press, Oxford, ISBN 0198504748.

Lesk A.M. 2002. *Introduction to Bioinformatics*. Oxford University Press, Oxford. ISBN 0199251967.

Lesk A.M. 2004. *Introduction to Protein Science.* Oxford University Press, Oxford, ISBN 9780199265114.

Letunic I. & Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23:** 127–128. https://doi.org/10.1093/bioinformatics/btl529

Majzlova K., Pukajova Z. & Janecek S. 2013. Tracing the evolution of the α-amylase subfamily GH13_36 covering the amylolytic enzymes intermediate between oligo-1,6-glucosidases and neopullulanases. *Carbohydr. Res.* **367:** 48–57. https://doi.org/10.1016/j.carres.2012.11.022

Markovic O. & Janecek S. 2001. Pectin degrading glycoside hydrolases of family 28: sequence-structural features, specificities and evolution. *Protein Eng.* **14:** 615–631. https://doi.org/10.1093/protein/14.9.615

Matsuura Y., Kusunoki M., Harada W. & Kakudo M. 1984. Structure and possible catalytic residues of Taka-amylase A. *J. Biochem.* **95:** 697–702. https://doi.org/10.1093/oxfordjournals.jbchem.a134659

Nelson K.E., Clayton R.A., Gill S.R., Gwinn M.L., Dodson R.J., Haft D.H., Hickey E.K., Peterson J.D., Nelson W.C., Ketchum K.A., McDonald L., Utterback T.R., Malek J.A., Linher K.D., Garrett M.M., Stewart A.M., Cotton M.D., Pratt M.S., Phillips C.A., Richardson D., Heidelberg J., Sutton G.G., Fleischmann R.D., Eisen J.A., White O., Salzberg S.L., Smith H.O., Venter J.C. & Fraser C.M. 1999. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* **399:** 323–329. https://doi.org/10.1038/20601

Ogasawara O., Kodama Y., Mashima J., Kosuge T. & Fujisawa T. 2020. DDBJ Database updates and computational infrastructure enhancement. *Nucleic Acids Res.* **48 (Database Issue 1):** D45–D50. https://doi.org/10.1093/nar/gkz982

Oslancova A. & Janecek S. 2002. Oligo-1,6-glucosidase and neopullulanase enzyme subfamilies from the α-amylase family defined by the fifth conserved sequence region. *Cell. Mol. Life. Sci.* **59:** 1945–1959. https://doi.org/10.1007/pl00012517

Parkhill J., Wren B.W., Thomson N.R., Titball R.W., Holden M.T., Prentice M.B., Sebaihia M., James K.D., Churcher C., Mungall K.L., Baker S., Basham D., Bentley S.D., Brooks K., Cerdeno-Tarraga A.M., Chillingworth T., Cronin A., Davies R.M., Davis P., Dougan G., Feltwell T., Hamlin N., Holroyd S., Jagels K., Karlyshev A.V., Leather S., Moule S., Oyston P.C., Quail M., Rutherford K., Simmonds M., Skelton J., Stevens K., Whitehead S. & Barrell B.G. 2001. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature* **413:** 523–527. https://doi.org/10.1038/35097083

Rost B., Yachdav G. & Liu J. 2004. The PredictProtein server. *Nucleic Acids Res.* **32 (Web Server issue):** W321–W326. https://doi.org/10.1093/nar/gkh377

Saitou N. & Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Sanger F., Air G.M., Barrell B.G., Brown N.L., Coulson A.R., Fiddes C.A., Hutchison C.A., Slocombe P.M. & Smith M. 1977. Nucleotide sequence of bacteriophage φX174 DNA. *Nature* 265: 687–695. https://doi.org/10.1038/265687a0

Sarian F.D., Janecek S., Pijning T., Ihsanawati, Nurachman Z., Radjasa O.K., Dijkhuizen L., Natalia D. & van der Maarel M.J. 2017. A new group of glycoside hydrolase family 13 α-amylases with an aberrant catalytic triad. *Sci. Rep.* **7:** 44230. https://doi.org/10.1038/srep44230

Sayers E.W., Beck J., Brister J.R., Bolton EE., Canese K., Comeau D.C., Funk K., Ketter A., Kim S., Kimchi A., Kitts P.A., Kuznetsov A., Lathrop S., Lu Z., McGarvey K., Madden T.L., Murphy T.D., O'Leary N., Phan L., Schneider V.A., Thibaud-Nissen F., Trawick B.W., Pruitt K.D. & Ostell J. 2020. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **48 (Database Issue 1):** D9–D16. https://doi.org/10.1093/nar/gkz899

Sayers E.W., Cavanaugh M., Clark K., Ostell J., Pruitt K.D. & Karsch-Mizrachi I. 2020 GenBank. *Nucleic Acids Res.* **48 (Database Issue 1):** D84–D86. https://doi.org/10.1093/nar/gkz956

Schwede T., Kopp J., Guex N. & Peitsch M.C. 2003. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31:** 3381-3385. https://doi.org/10.1093/nar/gkg520

Shatsky M., Nussinov R. & Wolfson H.J. 2004. A method for simultaneous alignment of multiple protein structures. *Proteins* **56:** 143–156. https://doi.rog/10.1002/prot.10628

Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D. & Higgins D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7:** 539. https://doi.org/10.1038/msb.2011.75

Tomb J.F., White O., Kerlavage A.R., Clayton R.A., Sutton G.G., Fleischmann R.D., Ketchum K.A., Klenk H.P., Gill S., Dougherty B.A., Nelson K., Quackenbush J., Zhou L., Kirkness E.F., Peterson S., Loftus B., Richardson D., Dodson R., Khalak H.G., Glodek A., McKenney K., Fitzegerald L.M., Lee N., Adams M.D., Hickey E.K., Berg D.E., Gocayne J.D., Utterback T.R., Peterson J.D., Kelley J.M., Cotton M.D., Weidman J.M., Fujii C., Bowman C., Watthey L., Wallin E., Hayes W.S., Borodovsky M., Karp P.D., Smith H.O., Fraser C.M. & Venter J.C. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547. https://doi.org/10.1038/41483

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47 (Database Issue 1):** D506–D515. https://doi.org/10.1093/nar/gky1049

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A., Gocayne J.D., Amanatides P., Ballew R.M., Huson D.H., Wortman J.R., Zhang Q., Kodira C.D., Zheng X.H., Chen L., Skupski M., Subramanian G., Thomas P.D., Zhang J., Gabor Miklos G.L., Nelson C., Broder S., Clark A.G., Nadeau J., McKusick V.A., Zinder N., Levine A.J., Roberts R.J., Simon M., Slayman C., Hunkapiller M., Bolanos R., Delcher A., Dew I., Fasulo D., Flanigan M., Florea L., Halpern A., Hannenhalli S., Kravitz S., Levy S., Mobarry C., Reinert K., Remington K., Abu-Threideh J., Beasley E., Biddick K., Bonazzi V., Brandon R., Cargill M., Chandramouliswaran I., Charlab R., Chaturvedi K., Deng Z., Di Francesco V., Dunn P., Eilbeck K., Evangelista C., Gabrielian A.E., Gan W., Ge W., Gong F., Gu Z., Guan P., Heiman T.J., Higgins M.E., Ji R.R., Ke Z., Ketchum K.A., Lai Z., Lei Y., Li Z., Li J., Liang Y., Lin X., Lu F., Merkulov G.V., Milshina N., Moore H.M., Naik A.K., Narayan V.A., Neelam B., Nusskern D., Rusch D.B., Salzberg S., Shao W., Shue B., Sun J., Wang Z., Wang A., Wang X., Wang J., Wei M., Wides R., Xiao C., Yan C., Yao A., Ye J., Zhan M., Zhang W., Zhang H., Zhao Q., Zheng L., Zhong F., Zhong W., Zhu S., Zhao S., Gilbert D., Baumhueter S., Spier G., Carter C., Cravchik A., Woodage T., Ali F., An H., Awe A., Baldwin D., Baden H., Barnstead M., Barrow I., Beeson K., Busam D., Carver A., Center A., Cheng M.L., Curry L., Danaher S., Davenport L., Desilets R., Dietz S., Dodson K., Doup L., Ferriera S., Garg N., Gluecksmann A., Hart B., Haynes J., Haynes C., Heiner C., Hladun S., Hostin D., Houck J., Howland T., Ibegwam C., Johnson J., Kalush F., Kline L., Koduru S., Love A., Mann F., May D., McCawley S., McIntosh T., McMullen I., Moy M., Moy L., Murphy B., Nelson K., Pfannkoch C., Pratts E., Puri V., Qureshi H., Reardon M., Rodriguez R., Rogers Y.H., Romblad D., Ruhfel B., Scott R., Sitter C., Smallwood M., Stewart E., Strong R., Suh E., Thomas R., Tint N.N., Tse S., Vech C., Wang G., Wetter J., Williams S., Williams M., Windsor S., Winn-Deen E., Wolfe K., Zaveri J., Zaveri K., Abril J.F., Guigó R., Campbell M.J., Sjolander K.V., Karlak B., Kejariwal A., Mi H., Lazareva B., Hatton T., Narechania A., Diemer K., Muruganujan A., Guo N., Sato S., Bafna V., Istrail S., Lippert R., Schwartz R., Walenz B., Yooseph S., Allen D., Basu A., Baxendale J., Blick L., Caminha M., Carnes-Stine J., Caulk P., Chiang Y.H., Coyne M., Dahlke C., Mays A., Dombroski M., Donnelly M., Ely D., Esparham S., Fosler C., Gire H., Glanowski S., Glasser K., Glodek A., Gorokhov M., Graham K., Gropman B., Harris M., Heil J., Henderson S., Hoover J., Jennings D., Jordan C., Jordan J., Kasha J., Kagan L., Kraft C., Levitsky A., Lewis M., Liu X., Lopez J., Ma D., Majoros W., McDaniel J., Murphy S., Newman M., Nguyen T., Nguyen N., Nodell M., Pan S., Peck J., Peterson M., Rowe W., Sanders R., Scott J., Simpson M., Smith T., Sprague A., Stockwell T., Turner R., Venter E., Wang M., Wen M., Wu D., Wu M., Xia A., Zandieh A. & Zhu X. 2001. The sequence of the human genome. *Science* **291:** 1304–1351. https://doi.org/10.1126/science.1058040

Venter J.C., Smith H.O. & Hood L. 1996. A new strategy for genome sequencing. *Nature* **381:** 364–366. https://doi.org/10.1038/381364a0

White O., Eisen J.A., Heidelberg J.F., Hickey E.K., Peterson J.D., Dodson R.J., Haft D.H., Gwinn M.L., Nelson W.C., Richardson D.L., Moffat K.S., Qin H., Jiang L., Pamphile W., Crosby M., Shen M., Vamathevan J.J., Lam P., McDonald L., Utterback T., Zalewski C., Makarova K.S., Aravind L., Daly M.J., Minton K.W., Fleischmann R.D., Ketchum K.A., Nelson K.E., Salzberg S., Smith H.O., Venter J.C. & Fraser C.M. 1999. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286:** 1571–1577. https://doi.org/10.1126/science.286.5444.1571

Woese C.R. & Fox G.E. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA* **74:** 5088–5090. https://doi.org/10.1073/pnas.74.11.5088

wwPDB consortium. 2019. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47 (Database Issue 1):** D520–D528. https://doi.org/10.1093/nar/gky949

Yachdav G., Kloppmann E., Kajan L., Hecht M., Goldberg T., Hamp T., Hönigschmid P., Schafferhans A., Roos M., Bernhofer M., Richter L., Ashkenazy H., Punta M., Schlessinger A., Bromberg Y., Schneider R., Vriend G., Sander C., Ben-Tal N. & Rost B. 2014. PredictProtein – an open resource for online prediction of protein structural and functional features. *Nucleic Acids Res.* **42 (Web Server issue):** W337–W343. https://doi.org/10.1093/nar/gku366

Zamocky M., Janecek S. & Koller F. 2000. Common phylogeny of catalase-peroxidases and ascorbate peroxidases. *Gene* **256:** 169–182. https://doi.org/10.1016/s0378-1119(00)00358-9

Zhang Z., Schäffer A.A., Miller W., Madden T.L., Lipman D.J., Koonin E.V. & Altschul S.F. 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26:** 3986–3990. https://doi.org/10.1093/nar/26.17.3986

BIOINFORMATIKA PROTEÍNOV

Autor: Doc. Ing. Štefan Janeček, DrSc.

Vydala: Univerzita sv. Cyrila a Metoda v Trnave, 2020

Vydanie: prvé.

Web: http://fpv.ucm.sk/images/ucebne_texty/Bioinformatika_proteinov.pdf

Počet strán: 106

ISBN 978-80-572-0085-7