MVAE Elektronikus Számolóközpont

Nagyhosszuságu aminosavszekvenciák közötti analógiák vizsgálata számitógéppel

Farkas András

Az alábbiakban nagyhosszuságu aminosavszekvenciák közötti speciális analógia meghatározásával foglalkozunk.

A probléma a MTA SZBK Biokémiai Intézete Enzimológiai Részlegénél végzett fehérjeszerkezet-vizsgálatok kapcsán merült fel. Biokémiai megfogalmazása Dévényi Tibortól származik.

A probléma a következő: adott két vagy több aminosavszekvencia, keressük azt a leghosszabb elméleti aminosavszekvenciát, amelynek elemeit mindegyik adott aminosavszekvencia tartalmazza mégpedig, ugyanabban a sorrendben, mint amilyen sorrendben az elméleti aminosavszekvenciában szerepelnek.

Az aminosavszekvenciák közötti speciális analógiát az ilyen elméleti aminosavszekvencia létezése jelenti. A probléma matematikai megfogalmazása két adott szekvencia esetén a következő:

Adottak az $\{a_i\} = a_1, a_2, \ldots, a_i, \ldots, a_n$

$\{b_i\} = b_1, b_2, \ldots, b_i, \ldots, b_m$ véges sorozatok.

A $\{c_k\} = c_1, c_2, \ldots, c_k, \ldots, c_\ell \quad \ell \leq \min(n,m)$

sorozatot $\{a_i\}$ és $\{b_i\}$ közös részsorozatának nevezzük, ha van olyan

$a_{i_1}, a_{i_2}, \ldots, a_{i_k}, \ldots, a_{i_\ell}$ részsorozata $\{a_i\}$ -nek,

továbbá van olyan

$b_{i_1}, b_{i_2}, \ldots, b_{i_k}, \ldots, b_{i_\ell}$ részsorozata $\{b_j\}$-nek,

hogy $c_k = a_{i_k} = b_{i_k}$ minden $1 \leq k \leq \ell$ esetén.

Feladat: $\{a_i\}$ és $\{b_j\}$ sorozatokhoz megadni a leghosszabb közös részsorozatot, tehát amelyre ℓ maximális.

Két adott sorozathoz a leghosszabb közös részsorozat meghatározására szolgáló algoritmus a következő:

Egy (n+1)x(m+1) -es mátrix (melynek sorait és oszlopait 0-tól sorszámozzuk) $\lambda_{i,j}$ elemeit a következő módon értelmezzük:

1.) Legyen a 0. sor és 0. oszlop minden eleme 0:

$$\lambda_{0,j} = 0 \qquad \lambda_{i,0} = 0$$
$$0 \leq j \leq m \qquad 0 \leq i \leq n$$

2.) Értelmezzük a további elemeket pl. oszlopfolytonosan szukcesszive:

$$\lambda_{i,j} = \begin{cases} 0 & \text{ha } a_i \neq b_j \\ \max\limits_{\substack{0 \leq p < i \\ 0 \leq q < j}} (\lambda_{p,q}) + 1 & \text{ha } a_i = b_j \end{cases} .$$

A maximális hosszuságu közös részsorozat, illetve részsorozatok a következő módon adódnak:

a maximális hosszuságu közös részsorozat hossza

$$\ell = \max_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} (\lambda_{i,j})$$

1.) Az utolsó eleme: $c_\ell = a_{i_\ell} = b_{i_\ell}$, ahol i_ℓ és j_ℓ az ℓ-el egyenlő $\lambda_{i,j}$ értékek közül valamelyiknek az indexei:

$$\lambda_{i_\ell, j_\ell} \in \left\{ \lambda_{i,j} : \lambda_{i,j} = \ell \right\}.$$

2.) H a k. eleme $c_k = a_{i_k} = b_{j_k}$, akkor a k-1. eleme

$$c_{k-1} = a_{i_{k-1}} = b_{j_{k-1}}$$

ahol i_{k-1} és j_{k-1} a k-1 értékkel egyenlő $\lambda_{i,j}$ elemek közül valamelyik olyannak az indexei, amelyre $i < i_k$ és $j < j_k$ fennáll.

$$\lambda_{i_{k-1}, j_{k-1}} \in \left\{ \lambda_{i,j} : \lambda_{i,j} = k-1, \quad \begin{matrix} i < i_k, \\ j < j_k \end{matrix} \right\}$$

Az algoritmus az összes lehetséges megoldást szolgáltatja.

A módszer általánositható 2 dimenzióról n dimenzióra, ilyenkor n darab sorozat közös részsorozatának meghatározására alkalmas. A módszer helyességére vonatkozó konstruktiv bizonyitás triviális.

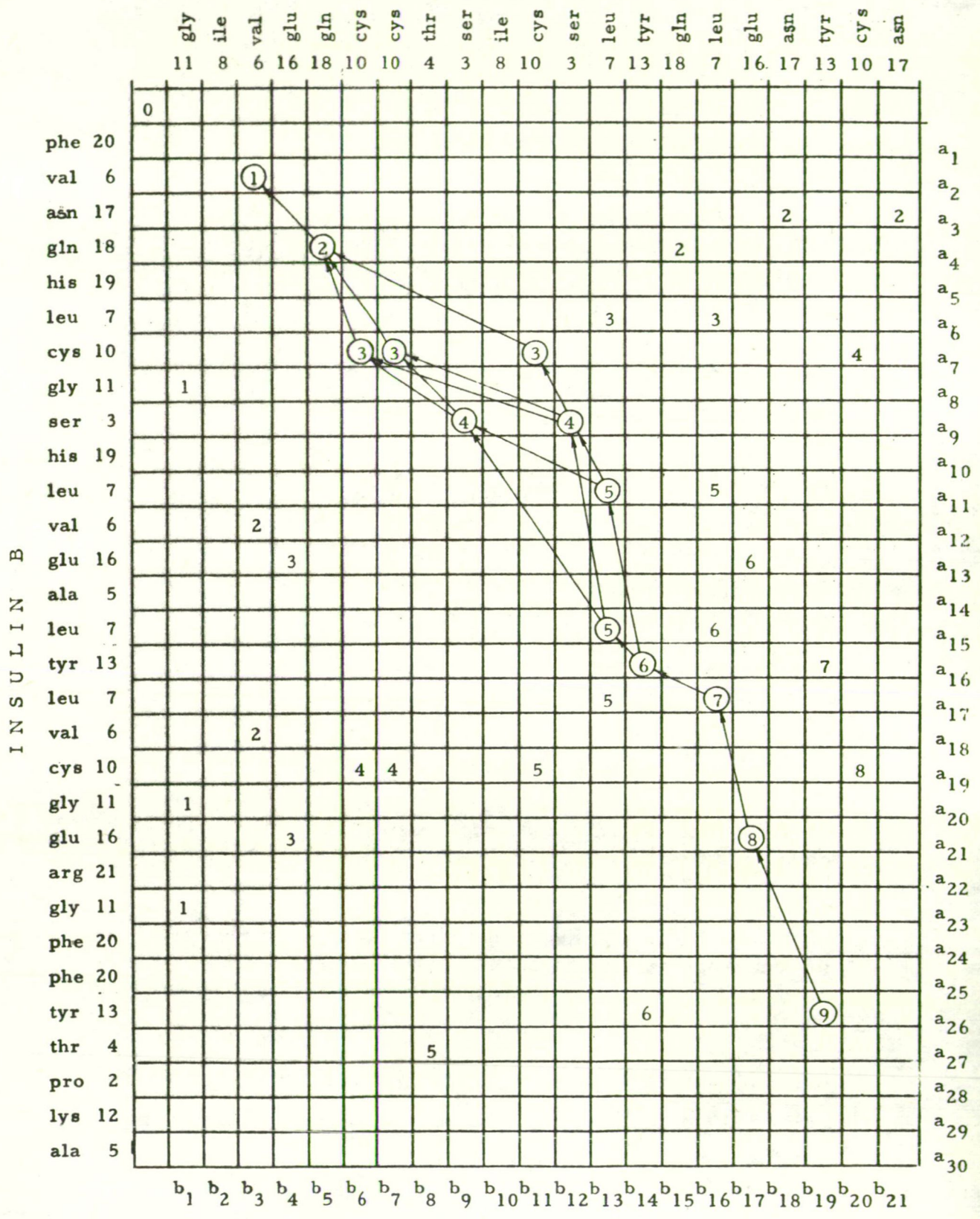
<u>Példa.</u> (ld. 234. old.)

A módszert különböző nagyhosszuságu aminosavszekvenciák leghosszabb közös részszekvenciájának meghatározására alkalmaztuk.

Itt a szarvasmarha májból származó 506 tagu glutaminsav dehidrogenáz, illetve a rák izomból származó 333 tagu gliceraldehid 3 foszfát dehidrogenáz aminosavszekvenciák esetén kapott eredményeinket ismertetjük.

A leghosszabb közös részszekvencia 151 tagunak adódott.

A következőkben ismertetjük a két aminosavszekvenciát és


INSULIN A
gly 11
ile 8
val 6
glu 16
gln 18
cys 10
cys 10
thr 4
ser 3
ile 8
cys 10
ser 3
leu 7
tyr 13
gln 18
leu 7
glu 16
asn 17
tyr 13
cys 10
asn 17
INSULIN B
phe 20
val 6
asn 17
gln 18
his 19
leu 7
cys 10
gly 11
ser 3
his 19
leu 7
val 6
glu 16
ala 5
leu 7
tyr 13
leu 7
val 6
cys 10
gly 11
glu 16
arg 21
gly 11
phe 20
phe 20
tyr 13
thr 4
pro 2
lys 12
ala 5
a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15 a16 a17 a18 a19 a20 a21 a22 a23 a24 a25 a26 a27 a28 a29 a30
b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b13 b14 b15 b16 b17 b18 b19 b20 b21

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
ALA	ASP	ARG	GLU	ASP	ASP	PRO	ASN	PHE	PHE	LYS	MET	VAL	GLU	GLY	PHE	PHE	ASP	ARG	GLY	ALA	SER	ILE	VAL	GLU
26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
ASP	LYS	LEU	VAL	GLU	ASP	LEU	LYS	THR	ARG	GLN	THR	GLN	GLU	GLN	LYS	ARG	ASN	ARG	VAL	ARG	GLY	ILE	LEU	ARG
51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
ALA	GLN	HIS	SER	HIS	GLN	ARG	THR	PRO	CYS	LYS	GLY	GLY	ILE	ARG	TYR	SER	THR	ASP	VAL	SER	VAL	ASP	GLU	VAL
76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
LYS	ALA	LEU	ALA	SER	LEU	MET	THR	TYR	LYS	CYS	ALA	VAL	VAL	ASP	VAL	PRO	PHE	GLY	GLY	ALA	LYS	ALA	GLY	VAL
101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125
LYS	ILE	ASN	PRO	LYS	ASN	TYR	THR	ASP	GLU	ASP	LEU	GLU	LYS	ILE	THR	ARG	THR	ARG	PHE	MET	GLU	LEU	THR	THR
126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
ALA	MET	GLU	LEU	ALA	LYS	LYS	GLY	PHE	ILE	GLY	PRO	GLY	LEU	ASP	VAL	PRO	ALA	PRO	ASN	MET	SER	THR	GLY	GLU
151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
ARG	GLU	MET	SER	TRP	ILE	ALA	ASP	THR	TYR	ALA	SER	THR	ILE	GLY	HIS	TYR	ASP	ILE	ASN	ALA	HIS	ALA	CYS	VAL
176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
THR	LYS	PRO	GLY	ILE	SER	GLN	GLY	GLY	ILE	HIS	GLY	ARG	ILE	SER	ALA	THR	GLY	ARG	GLY	VAL	PHE	GLY	HIS	ILE
201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225
GLU	ASN	PHE	ILE	GLU	ASN	ALA	SER	TYR	MET	SER	ILE	LEU	GLY	MET	THR	PRO	GLY	PHE	GLY	ASP	LYS	THR	PHE	ALA
226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250
VAL	GLN	GLY	PHE	GLY	ASN	VAL	GLY	LEU	HIS	SER	MET	ARG	TYR	LEU	HIS	ARG	PHE	GLY	ALA	LYS	CYS	VAL	ALA	VAL
251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275
GLY	GLU	SER	ASP	GLY	SER	ILE	TRP	ASN	PRO	ASP	GLY	ILE	ASP	PRO	LYS	GLU	LEU	GLU	ASP	PHE	LYS	LEU	GLN	HIS
276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300
GLY	THR	ILE	LEU	GLY	PHE	PRO	LYS	ALA	LYS	ILE	TYR	GLU	GLY	SER	ILE	LEU	GLU	VAL	ASP	GLY	ASP	ILE	LEU	ILE
301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325
PRO	ALA	ALA	SER	GLU	LYS	GLN	LEU	THR	LYS	SER	ASN	ALA	PRO	ARG	VAL	LYS	ALA	LYS	ILE	ILE	ALA	GLU	GLY	ALA
326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350
ASN	GLY	PRO	THR	THR	PRO	GLY	ALA	ASP	LYS	ILE	PHE	LEU	GLU	ARG	ILE	ILE	LYS	PRO	CYS	ASN	HIS	VAL	LEU	SER
351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375
LEU	SER	PHE	PRO	ILE	ARG	ARG	ASP	ASP	GLY	SER	THR	GLU	VAL	ILE	GLU	GLY	TYR	ARG	ILE	GLY	MET	VAL	ILE	PRO
376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400
ASP	LEU	TYR	LEU	ASN	ALA	GLY	GLY	VAL	THR	VAL	SER	TYR	PHE	GLY	LEU	LYS	ASN	LEU	ASN	HIS	VAL	SER	TYR	GLY
401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425
ARG	LEU	THR	PHE	LYS	TYR	GLU	ARG	ASP	SER	ASN	TYR	HIS	LEU	LEU	MET	SER	VAL	GLN	GLU	SER	LEU	GLU	ARG	LYS
426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450
PHE	GLY	LYS	HIS	GLY	GLY	THR	ILE	PRO	ILE	VAL	PRO	THR	ALA	GLU	PHE	GLN	ASP	ARG	ILE	SER	GLY	ALA	SER	GLU
451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475
LYS	ASP	ILE	VAL	HIS	SER	GLY	LEU	ALA	TYR	THR	MET	GLU	ARG	SER	ALA	ARG	GLN	ILE	MET	ARG	THR	ALA	MET	LYS
476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500
TYR	ASN	LEU	GLY	LEU	ASP	LEU	ARG	THR	ALA	ALA	TYR	VAL	ASN	ALA	ILE	GLU	LYS	VAL	PHE	ARG	VAL	TYR	ASN	GLU
501	502	503	504	505	506																			
ALA	GLY	VAL	THR	PHE	THR																			

Glutaminsav dehidrogenáz.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
SER	LYS	ILE	GLY	ILE	ASP	GLY	PHE	GLY	ARG	ILE	GLY	ARG	LEU	VAL	LEU	ARG	ALA	ALA	LEU	SER	CYS	GLY	ALA	GLN

26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
VAL	VAL	ALA	VAL	ASN	ASP	PRO	PHE	ILE	ALA	LEU	GLU	TYR	MET	VAL	TYR	MET	PHE	LYS	TYR	ASP	SER	THR	HIS	GLY

51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
VAL	PHE	LYS	GLY	GLU	VAL	LYS	MET	GLU	ASP	GLY	ALA	LEU	VAL	VAL	ASP	GLY	LYS	LYS	ILE	THR	VAL	PHE	ASN	GLU

76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
MET	LYS	PRO	GLU	ASN	ILE	PRO	TRP	SER	LYS	ALA	GLY	ALA	GLU	TYR	ILE	VAL	GLU	SER	THR	GLY	VAL	PHE	THR	THR

101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125
ILE	GLU	LYS	ALA	SER	ALA	HIS	PHE	LYS	GLY	GLY	ALA	LYS	LYS	VAL	VAL	ILE	SER	ALA	PRO	SER	ALA	ASP	ALA	PRO

126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150
MET	PHE	VAL	CYS	GLY	VAL	ASN	LEU	GLU	LYS	TYR	SER	LYS	ASP	MET	THR	VAL	VAL	SER	ASN	ALA	SER	CYS	THR	THR

151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169	170	171	172	173	174	175
ASN	CYS	LEU	ALA	PRO	VAL	ALA	LYS	VAL	LEU	HIS	GLU	ASN	PHE	GLU	ILE	VAL	GLU	GLY	LEU	MET	THR	THR	VAL	HIS

176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
ALA	VAL	THR	ALA	THR	GLN	LYS	THR	VAL	ASP	GLY	PRO	SER	ALA	LYS	ASP	TRP	ARG	GLY	GLY	ARG	GLY	ALA	ALA	GLN

201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225
ASN	ILE	ILE	PRO	SER	SER	THR	GLY	ALA	ALA	LYS	ALA	VAL	GLY	LYS	VAL	ILE	PRO	GLU	LEU	ASP	GLY	LYS	LEU	THR

226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250
GLY	MET	ALA	PHE	ARG	VAL	PRO	THR	PRO	ASP	VAL	SER	VAL	VAL	ASP	LEU	THR	VAL	ARG	LEU	GLY	LYS	GLU	CYS	SER

251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275
TYR	ASP	ASP	ILE	LYS	ALA	ALA	MET	LYS	THR	ALA	SER	GLU	GLY	PRO	LEU	GLN	GLY	PHE	LEU	GLY	TYR	THR	GLU	ASP

276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300
ASP	VAL	VAL	SER	SER	ASP	PHE	ILE	GLY	ASP	ASN	ARG	SER	SER	ILE	PHE	ASP	ALA	LYS	ALA	GLY	ILE	GLN	LEU	SER

301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325
LYS	THR	PHE	VAL	LYS	VAL	VAL	SER	TRP	TYR	ASP	ASN	GLU	PHE	GLY	TYR	SER	GLN	ARG	VAL	ILE	ASP	LEU	LEU	LYS

326	327	328	329	330	331	332	333
HIS	MET	GLN	LYS	VAL	ASP	SER	ALA

Gliceraldehid 3 foszfát dehidrogenáz.

a megfelelő aminosavak aláhuzásával jelöljük azokat az aminosavakat, amelyek a leghosszabb közös részszekvenciába tartoznak.

A lehetséges megoldások közül egy "bal-alsó" megoldást választottunk.

Az eredményként adódó elméleti aminosavszekvenciát meg kell vizsgálnunk abból a szempontból, hogy hossza "kirivóan" nagy-e, azaz valószinütlen esemény-e az, hogy a megvizsgált aminosavszekvenciákkal azonos hosszuságu és aminosavösszetételü, de véletlen aminosav-sorrendü szekvenciákban legalább az eredményben szereplő leghosszabb közös aminosav--szekvenciahossz adódik.

Ha valóban valószinütlen esemény, akkor szükséges a "kirivó" hossz jelenségének biokémiai interpretációja.

A valószinüséget sztochasztikus szimulációval határozhatjuk meg, tekintettel arra, hogy a valószinüség explicit alakban való megadása problematikus.