Using Amino Acid Analysis for Protein Identification

Sally Bee and Elisabeth Gasteiger*
Biochrom Ltd, Cambridge, UK
*Swiss Institute of Bioinformatics, Geneva, Switzerland

Complete hydrolysis of a protein to its component amino acids allows automated separation and quantification at the pmole to nanomole level. Although sequence information is lost, compositional data provides a protein specific fingerprint for identification purposes.

Compositional data of peptides and proteins can be determined using an amino acid analyser.

If the identity of a protein is unknown, the compositional data of the protein can be used in searching protein databases.

A programme that provides this function is the AA CompIdent tool from the Swiss Institute of Bioinformatics. This database can be found at www.expasy.ch/tools/aacomp.

The tool can be used to identify a protein from its amino acid composition. Swiss-Prot and/or TrEMBL databases are used to search for amino acid compositions that are closest to the amino acid composition given (1)

As an experiment, the human insulin beta chain was chosen as the test protein and the amino acid composition of the this protein was input into the AA CompIdent tool (see tables 1 and 2).

Choosing a protein of known composition allowed us to test the accuracy of the AA CompIdent tool, and determine whether using the amino acid composition would allow us to identify the correct protein from the database.

The programme matches the percentage empirically measured amino acid composition of the protein against the theoretical percentage amino acid composition of proteins. Results tables are emailed to the user (see Table 3 for the returned table). The score represents the degree of difference between the composition of the

Species searched	HOMO SAPIENS
pI	7.00
Mw	5000

Amino Acid	% composition
Asp	0.00
Gln	3.30
Gly	10.00
Pro	0.00
Val	10.00
Ile	0.00
Phe	0.00
Asp	3.30
Ser	3.30
Thr	6.70
Tyr	6.70
Met	0.00
Leu	13.30
Lys	3.30
Glu	6.70
His	6.70
Ala	3.30
Arg	3.30
Cys	6.70
Trp	0.00

Table 1 and 2. Data input into the AA CompIdent tool

unknown protein and a protein in the database. The score is calculated for each database entry by the sum of the squared difference between the percentage amino acid composition for all amino acids of the unknown protein and the database entry.

Rank	Score	Protein	Description
1	11	INS HUMAN	Insulin B chain
2	112	ZDH4 HUMAN	Zn finger DHHC domain containing
3	136	CIW1_HUMAN	Potassium channel subfamily K member 1
4	140	CTE2_HUMAN	Protein C20orf142
5	143	2B14_HUMAN	HLA class II histocompatibility antigen
6	145	PLCC_HUMAN	1-acyl-sn-glycerol-3-phosphate
7	150	DER7_HUMAN	31680 Dermal papilla derived protein 7.
8	151	2B32_HUMAN	HLA class II histocompatibility antigen
9	152	SPR1_HUMAN	Sphingosylphosphorylcholine receptor
10	152	C151_HUMAN	Platelet-endothelial tetraspan antigen 3
11	154	GP42_HUMAN	Probable G protein-coupled receptor
12	154	ICE2_HUMAN	Caspase-2 subunit p18.
13	156	2B31_HUMAN	HLA class II histocompatibility antigen
14	157	MYPR_HUMAN	Myelin proteolipid protein (PLP)
15	158	2B18_HUMAN	HLA class II histocompatibility antigen
16	160	GP97_HUMAN	Probable G protein-coupled receptor 97
17	160	B4G4_HUMAN	Beta-1,4-galactosyltransferase 4
18	161	ZDHF_HUMAN	Zinc finger DHHC domain containing
19	163	GP41_HUMAN	Putative G protein-coupled receptor
20	163	ADR1_HUMAN	Adiponectin receptor protein

Table 3. Output emailed to user

All proteins in the database are ranked according to their score, from lowest (best match) to highest (worst match), see table 3.

The AACompident Tool ranks Insulin Beta Chain as the highest match (marked with a star), which gives us confidence in using this tool to identify unknown proteins using amino acid composition.

Using the Biochrom 30 Amino Acid Analyser to identify proteins provides an additional application for the instrument.

Reference

1. **Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M.**

   The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003 Nucleic Acids Res. 31:365-370 (2003).