COPid: Composition Based Protein Identification

User's Manual

G. P. S. Raghava

[1]Bioinformatics Centre. Institute of Microbial Technology. Sector 39-A,
Chandigarh, INDIA
Fax: +91-172-2690632 or 2690585. Phone: +91-172-2690557 or 2690225
*Email:raghava@imtech.res.in

COPid: An introduction

A large number of highly sophisticated alignment methods like BLAST and FASTA have been developed to perform sequence similarity search. No body can question about their role in function prediction but they fail in the absence of significant similarity with any annotated protein. It has been observed that few class of proteins maintain similar/identical residue composition despite poor sequence similarity among themselves. In this situation simple alignment free or composition-based similarity technique can be more informative than sophisticated alignment methods in prediction and classification.

In order to exploit full potential of composition based method, we developed a comprehensive software package COPid. This package has three main modules; i) similarity search; ii) composition calculation and iii) analysis. The search module allows searching of similar sequences based on either amino acid or dipeptide composition of whole or N/C-terminus of a protein. The composition calculation module allows computation of residue composition. The analysis module allows one to perform; i) comparison of composition of two group of proteins; ii) creation of phylo-genetic distance matrix based on protein composition and iii) generation of patterns for using popular machine learning techniques like ANN and SVM.

All the tools are made available in the form of webserver that can be accessed at http://www.imtech.res.in/raghava/copid/. It is developed and launched on SUN server 420R under Solaris environment by using public domain software package Apache. All web pages are scripted in Hyper Text Markup Language (HTML). Some JavaScripts are also embedded in the web pages to make it more user friendly and interactive. Back end programming is done in CGI-perl and C. In addition COPid has been also mirrored at University of Arkansas for Medical Sciences, Little Rock, USA on Apple Macintosh Bioinformatics cluster (http://bioinformatics.uams.edu/raghava/copid/).

How to use COPid

1. Home page

Home page of the server contains three sections (Figure 1):

1.1. Top menu: It contains link to relevant information at the top. The menu has following options:

1.1.1. **Home:** directs to the home page of the server.

1.1.2. **Search Algorithms:** provides the algorithms by which searching is being done.

1.1.3. **Help:** this option is linked to the page that contain detailed information about the various options of web server.

1.1.4. **Team:** directs to information of the persons involved in developing the server.

1.1.5. **Contact:** clicking this link can access address and E-mail of the person to be contacted in case of some problem with using server or any other information.

1.2. Hyperlink to the different modules of server:

1.2.1. **Search:**

1.2.1.1. **against standard databases:** This provides the option of searching the proteins of similar composition (either amino acid or dipeptide) in standard databases like swissprot and PDB. The searching can be done on the basis of composition of either whole or N-terminal or C-terminal of proteins. The method of searching is described in section 2.2

1.2.1.2. **against mydatabase:** In this module, user can upload his/her protein sequences, which will be used as database during searching. The method of searching is described in section 2.2

1.2.2. **Composition:** It basically calculates the composition (amino acid or diepeptide) of either terminal or whole protein as specified by the user. If more than one sequence are given then average composition will also be calculated. The process of using this module is described in section 2.3

1.2.3. **Analysis:** By using this module, distance matrix can be constructed for OC and WebPhylip (section 2.4) amino acid or dipeptide composition of two group of sequences can be compared (section 2.5) and amino acid or dipeptide patterns for SNNS or SVM or Timbl can be constructed (section 2.6).

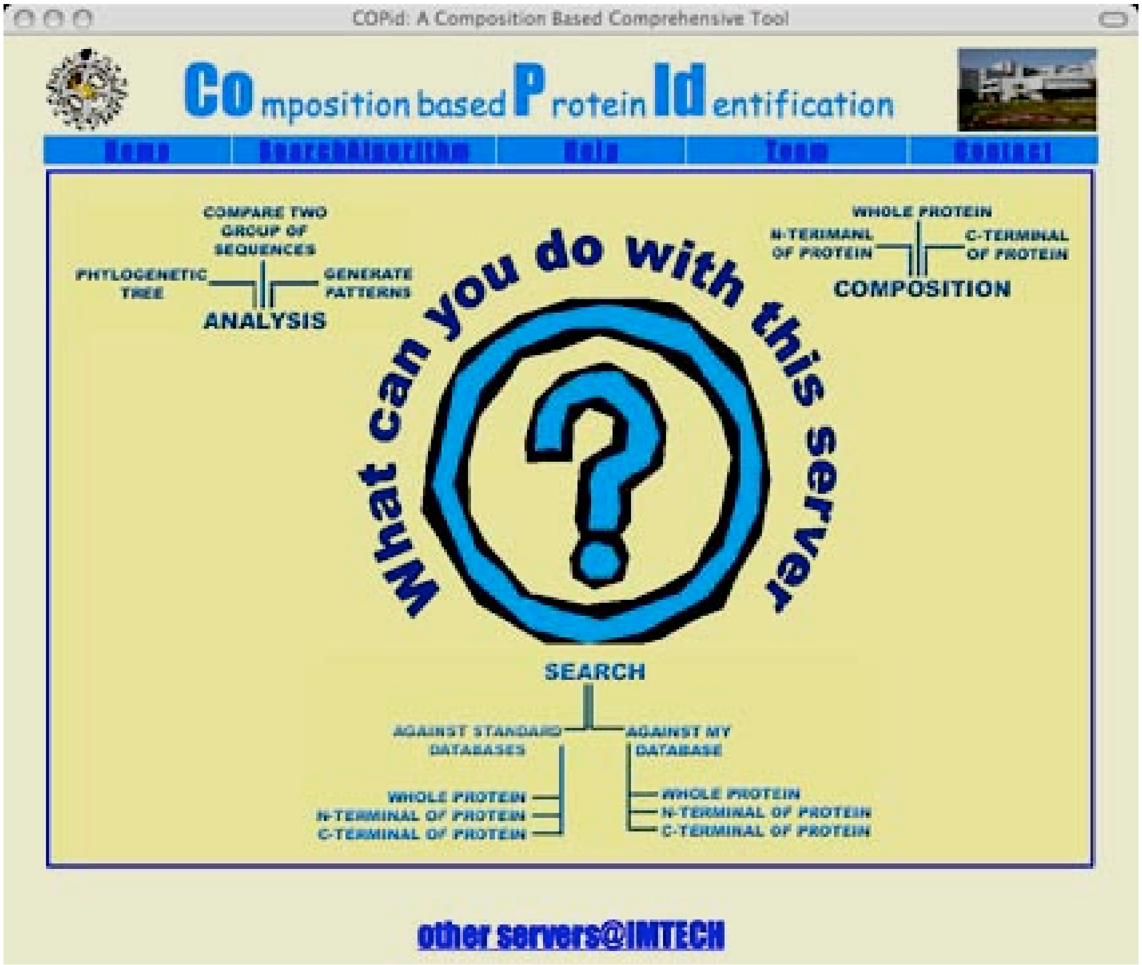


Figure 1: Home page of COPid web server.

1.3. **Footer hyperlink:** By clicking this, one can navigate to the page enlisting all the servers developed by Dr. GPS Raghava group and informations about the bioinformatics center at Institute of Microbial Technology, Chandigarh (160036), India.

Except the hyperlink to the different modules of server, both top menu and footer is present in all pages including the result.

2. Instructions to use the server:

The sequence submission form is a web interface wherein users can paste their sequence(s), select among the choices provided, composition/parameter of their choice and submit it to the server that returns the result of this query.

2.1. The fields that are optional and few general instructions are as follows:

2.1.1.**Name of job:** The name of job is purely an optional one. It can be any alphanumeric character.

2.1.2.**Email address:** It is also an optional field. Since each job is placed in the que, so depending upon the length of que getting search result can take quite a bit of time. To avoid this, email address can be provided, on which information of job completion will be intimated.

2.1.3.**Protein sequence:** It must be in single letter code, that can be either pasted or uploaded from a local file. All the non standard characters like [*&^%$@#!()_+~=;'",<>?.\|} are ignored from the sequence. Although the server uses ReadSeq routine to parse the input, but still it is requested to use the sequences that are strictly in FASTA format.

2.2.Searching proteins of similar amino acid or dipeptide composition (Figure 2a):

2.2.1.Click appropriate hyperlink under search on home page or unfold the search menu at left hand side of any submission form, then click appropriate option.

2.2.2.Enter name of job (optional).

2.2.3.Enter e-mail id on which you want to get information regarding completion of job (optional).

2.2.4.Submit one or more sequences in FASTA format by either pasting in the text box or uploading the file.

2.2.5.Select the database (for search against standard database) or upload the sequences which will be used as database.

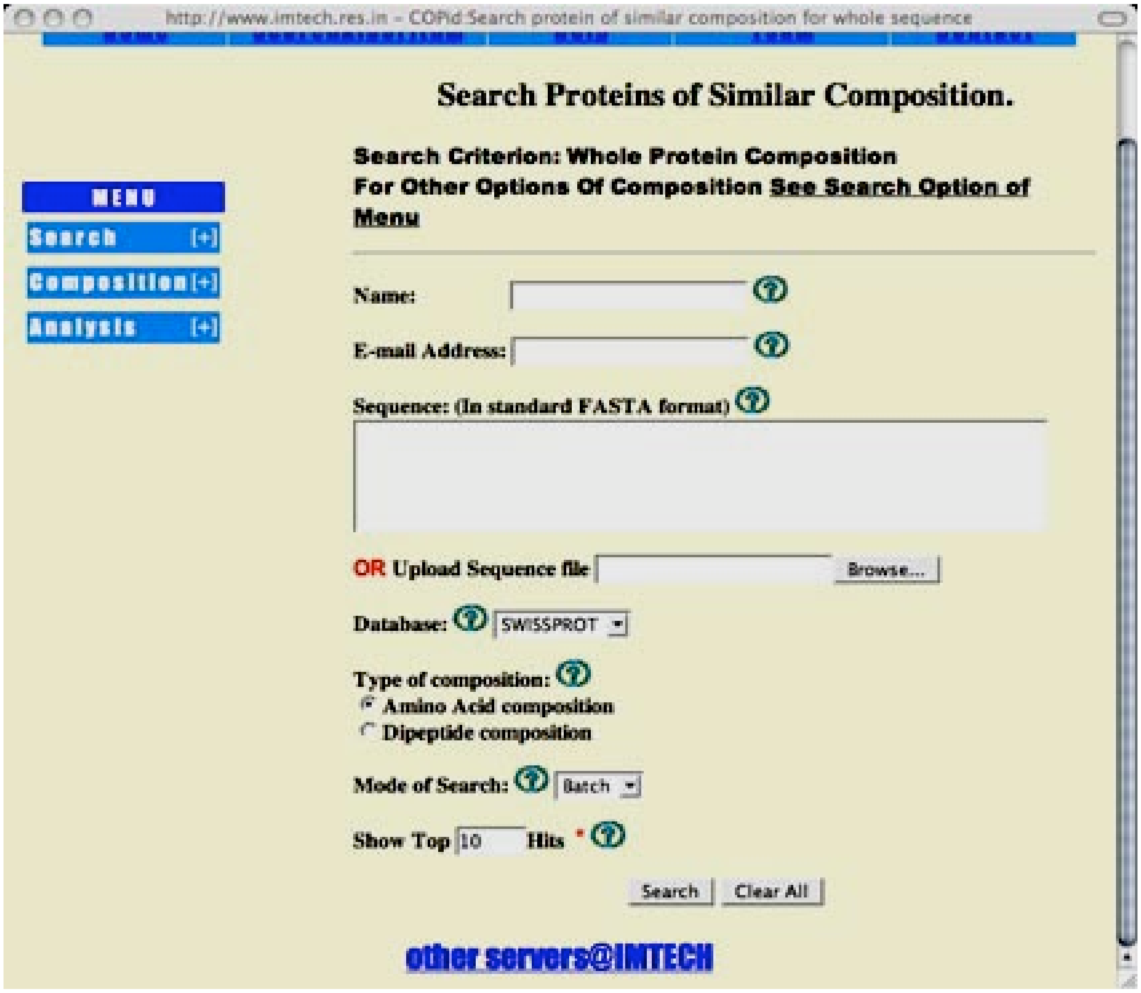
2.2.6.Mark the composition which will be used for searching.

2.2.7.Select between batch and mean mode of searching.

2.2.8.Specifiy the number of residues, whose composition will be used for searching (for terminal composition based search only).

2.2.9.Give the number of top hits reported after the searching.

2.2.10. Click submit to start searching or reset to clear all data in the form.


http://www.imtech.res.in - COPid:Search protein of similar composition for whole sequence
MENU
Search [+]
Composition [+]
Analysis [+]
Browse...
Search
Clear All
other servers@IMTECH

Search Proteins of Similar Composition.

Search Criterion: Whole Protein Composition
For Other Options Of Composition See Search Option of Menu

Name:

E-mail Address:

Sequence: (In standard FASTA format)

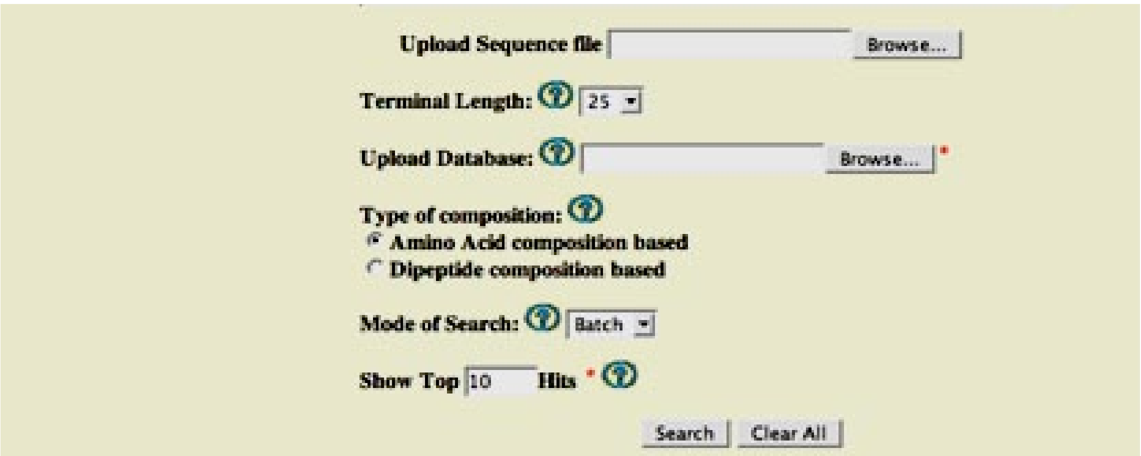
OR Upload Sequence file

Database: SWISSPROT

Type of composition:
● Amino Acid composition
○ Dipeptide composition

Mode of Search: Batch

Show Top 10 Hits *



Upload Sequence file

Terminal Length: 25

Upload Database: *

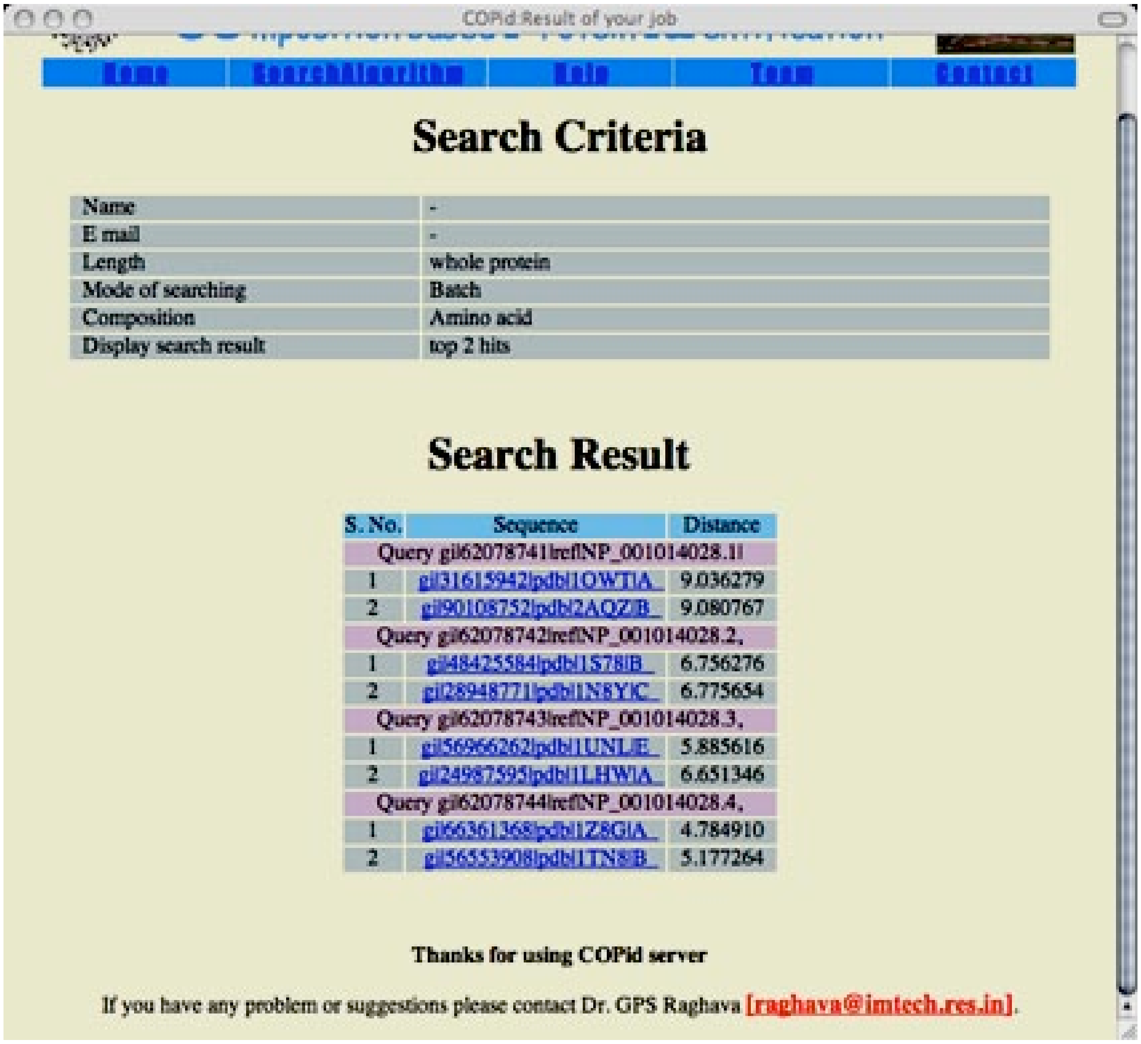
Type of composition:
● Amino Acid composition based
○ Dipeptide composition based

Mode of Search: Batch

Show Top 10 Hits *

Figure 2a: Submission form in which sequence(s) can be submitted to search for proteins with similar amino acid/dipeptide composition. First figure shows the submission form of searching against standard databases (PDB) with whole protein composition. Second one searche with terminal composition in user specified database.

The search result (figure 2b) will a list of specified number of proteins listed in the ascending order of Euclidian distance between query and database proteins.



Search Criteria

Name	-
E mail	-
Length	whole protein
Mode of searching	Batch
Composition	Amino acid
Display search result	top 2 hits

Search Result

S. No.	Sequence	Distance
Query gi|62078741|ref|NP_001014028.1|		
1	gi|31615942|pdb|1OWT|A_	9.036279
2	gi|90108752|pdb|2AQZ|B_	9.080767
Query gi|62078742|ref|NP_001014028.2,		
1	gi|48425584|pdb|1S78|B_	6.756276
2	gi|28948771|pdb|1N8Y|C_	6.775654
Query gi|62078743|ref|NP_001014028.3,		
1	gi|56966262|pdb|1UNL|E_	5.885616
2	gi|24987595|pdb|1LHW|A_	6.651346
Query gi|62078744|ref|NP_001014028.4,		
1	gi|66361368|pdb|1Z8G|A_	4.784910
2	gi|56553908|pdb|1TN8|B_	5.177264

Thanks for using COPid server

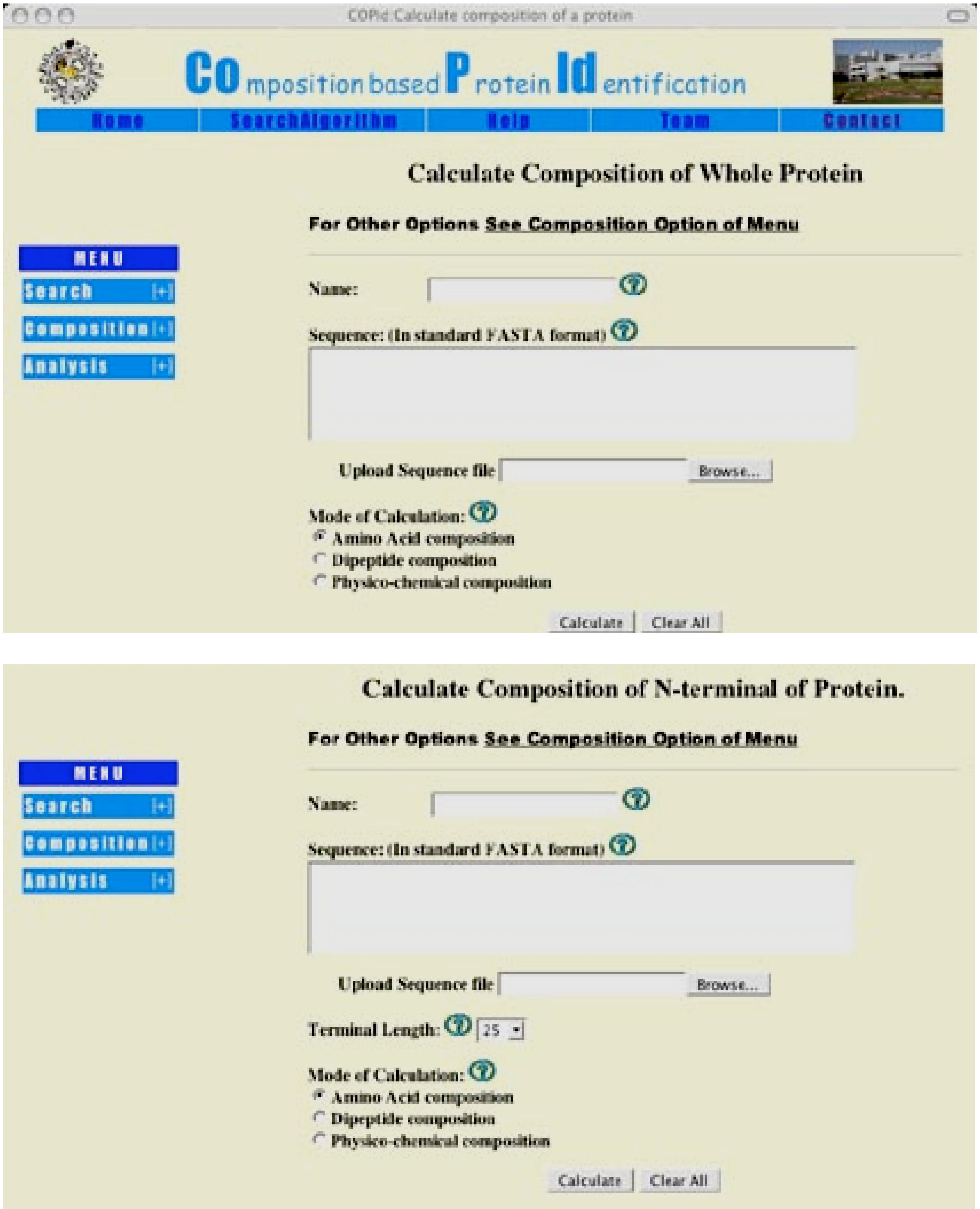
If you have any problem or suggestions please contact Dr. GPS Raghava [raghava@imtech.res.in].

Figure 2b: Composition based search result. In this search three proteins are used as query and only top two hits per query proteins are requested to display.

2.3.Calculation of amino acid or dipeptide composition (Figure 3):

2.3.1.Click appropriate hyperlink under composition on home page or unfold the composition menu at left hand side of any submission form, then click appropriate option.
2.3.2.Enter name of job (optional).
2.3.3.Submit one or more sequences in FASTA format by either pasting in the text box or uploading the file.
2.3.4.Select the number of residues up to which calculation is being done (for terminal composition only).

2.3.5.Choose the composition that user wish to calculate.
2.3.6.Click submit to calculate composition or reset to clear all data in the form.



COPid:Calculate composition of a protein

COmposition based Protein IDentification

Calculate Composition of Whole Protein

For Other Options See Composition Option of Menu

Name:

Sequence: (In standard FASTA format)

Upload Sequence file

Mode of Calculation:
● Amino Acid composition
○ Dipeptide composition
○ Physico-chemical composition

Calculate Composition of N-terminal of Protein.

For Other Options See Composition Option of Menu

Name:

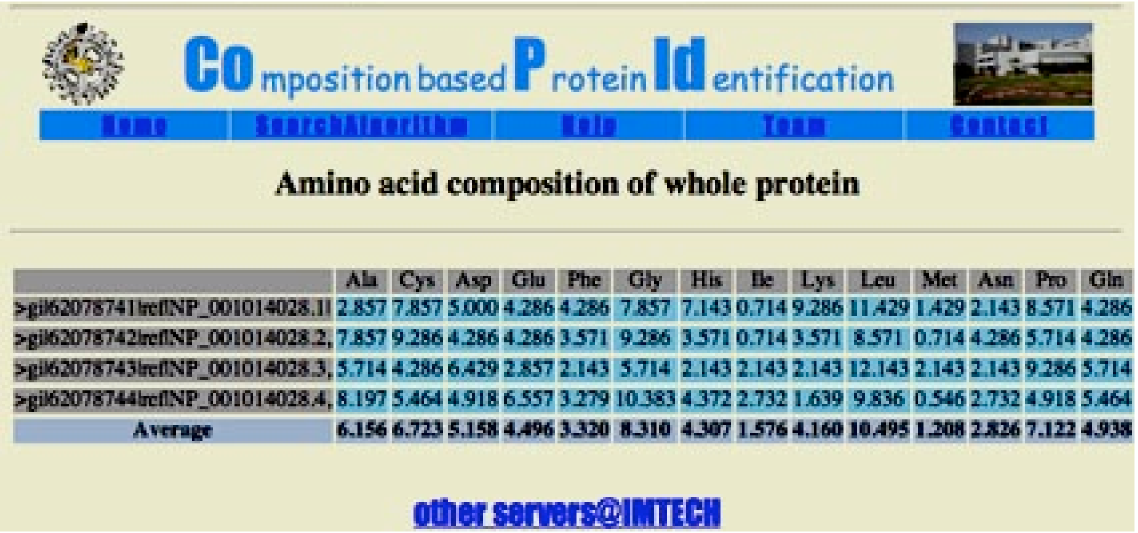
Sequence: (In standard FASTA format)

Upload Sequence file

Terminal Length: 25

Mode of Calculation:
● Amino Acid composition
○ Dipeptide composition
○ Physico-chemical composition

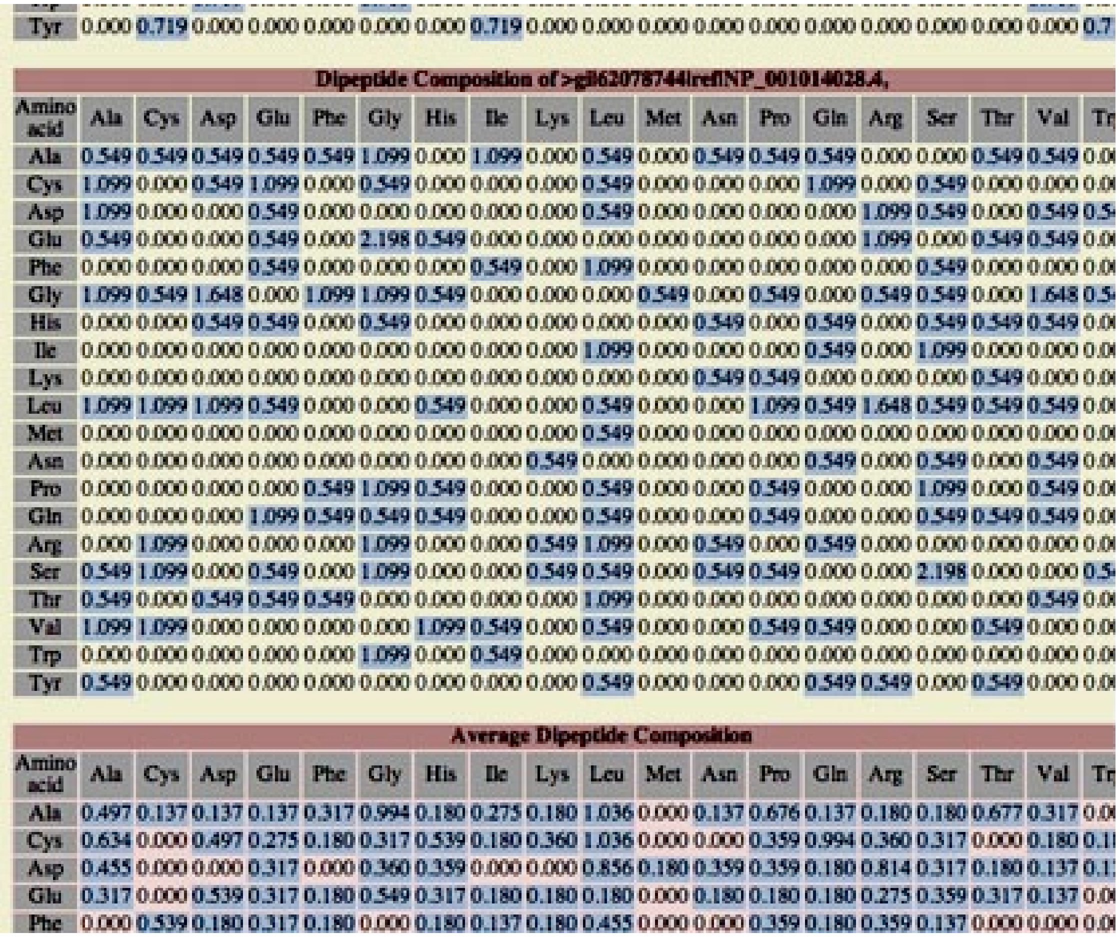
Figure 3a: Submission form for calculation of amino acid/dipeptide composition of proteins. Upper form will be used in case of whole protein composition while lower form will be used for composition of terminal composition.



Amino acid composition of whole protein

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln
>gi|62078741|ref|NP_001014028.1|	2.857	7.857	5.000	4.286	4.286	7.857	7.143	0.714	9.286	11.429	1.429	2.143	8.571	4.286
>gi|62078742|ref|NP_001014028.2,	7.857	9.286	4.286	4.286	3.571	9.286	3.571	0.714	3.571	8.571	0.714	4.286	5.714	4.286
>gi|62078743|ref|NP_001014028.3,	5.714	4.286	6.429	2.857	2.143	5.714	2.143	2.143	2.143	12.143	2.143	2.143	9.286	5.714
>gi|62078744|ref|NP_001014028.4,	8.197	5.464	4.918	6.557	3.279	10.383	4.372	2.732	1.639	9.836	0.546	2.732	4.918	5.464
Average	6.156	6.723	5.158	4.496	3.320	8.310	4.307	1.576	4.160	10.495	1.208	2.826	7.122	4.938

other servers@IMTECH



Tyr	0.000	0.719	0.000	0.000	0.000	0.000	0.000	0.719	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.7

Dipeptide Composition of >gi|62078744|ref|NP_001014028.4,

Amino acid	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Tr
Ala	0.549	0.549	0.549	0.549	0.549	1.099	0.000	1.099	0.000	0.549	0.000	0.549	0.549	0.549	0.000	0.000	0.549	0.549	0.0
Cys	1.099	0.000	0.549	1.099	0.000	0.549	0.000	0.000	0.000	0.549	0.000	0.000	0.000	1.099	0.000	0.549	0.000	0.000	0.0
Asp	1.099	0.000	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.000	1.099	0.549	0.000	0.549	0.5
Glu	0.549	0.000	0.000	0.549	0.000	2.198	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.099	0.000	0.549	0.549	0.0
Phe	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.549	0.000	1.099	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.0
Gly	1.099	0.549	1.648	0.000	1.099	1.099	0.549	0.000	0.000	0.000	0.549	0.000	0.549	0.000	0.549	0.549	0.000	1.648	0.5
His	0.000	0.000	0.549	0.549	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.549	0.000	0.549	0.549	0.549	0.0
Ile	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.099	0.000	0.000	0.000	0.549	0.000	1.099	0.000	0.000	0.0
Lys	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.549	0.000	0.000	0.000	0.549	0.000	0.0
Leu	1.099	1.099	1.099	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.000	0.000	1.099	0.549	1.648	0.549	0.549	0.549	0.0
Met	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
Asn	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.000	0.549	0.000	0.549	0.000	0.549	0.0
Pro	0.000	0.000	0.000	0.000	0.549	1.099	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.000	0.000	1.099	0.000	0.549	0.0
Gln	0.000	0.000	0.000	1.099	0.549	0.549	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.549	0.549	0.0
Arg	0.000	1.099	0.000	0.000	0.000	1.099	0.000	0.000	0.549	1.099	0.000	0.549	0.000	0.549	0.000	0.000	0.000	0.000	0.0
Ser	0.549	1.099	0.000	0.549	0.000	1.099	0.000	0.000	0.549	0.549	0.000	0.549	0.549	0.000	0.000	2.198	0.000	0.000	0.5
Thr	0.549	0.000	0.549	0.549	0.549	0.000	0.000	0.000	0.000	1.099	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.0
Val	1.099	1.099	0.000	0.000	0.000	0.000	1.099	0.549	0.000	0.549	0.000	0.000	0.549	0.549	0.000	0.000	0.549	0.000	0.0
Trp	0.000	0.000	0.000	0.000	0.000	1.099	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
Tyr	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.549	0.549	0.000	0.549	0.000	0.0

Average Dipeptide Composition

Amino acid	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Tr
Ala	0.497	0.137	0.137	0.137	0.317	0.994	0.180	0.275	0.180	1.036	0.000	0.137	0.676	0.137	0.180	0.180	0.677	0.317	0.0
Cys	0.634	0.000	0.497	0.275	0.180	0.317	0.539	0.180	0.360	1.036	0.000	0.000	0.359	0.994	0.360	0.317	0.000	0.180	0.1
Asp	0.455	0.000	0.000	0.317	0.000	0.360	0.359	0.000	0.000	0.856	0.180	0.359	0.359	0.180	0.814	0.317	0.180	0.137	0.1
Glu	0.317	0.000	0.539	0.317	0.180	0.549	0.317	0.180	0.180	0.180	0.000	0.180	0.180	0.180	0.275	0.359	0.317	0.137	0.0
Phe	0.000	0.539	0.180	0.317	0.180	0.000	0.180	0.137	0.180	0.455	0.000	0.000	0.359	0.180	0.359	0.137	0.000	0.000	0.0

Figure 3b: The composition calculated by COPid web-server. In case of amino acid composition, each row shows the composition of a protein (upper panel). While the last row contains the average composition if more than one protein is submitted. Dipeptide composition is displayed in tabular form (lower panel). First dipeptide composition of each protein is displayed one by one followed by the their average.

2.4.Creation of distance matrix for construction of phylogenetic tree (Figure 4):

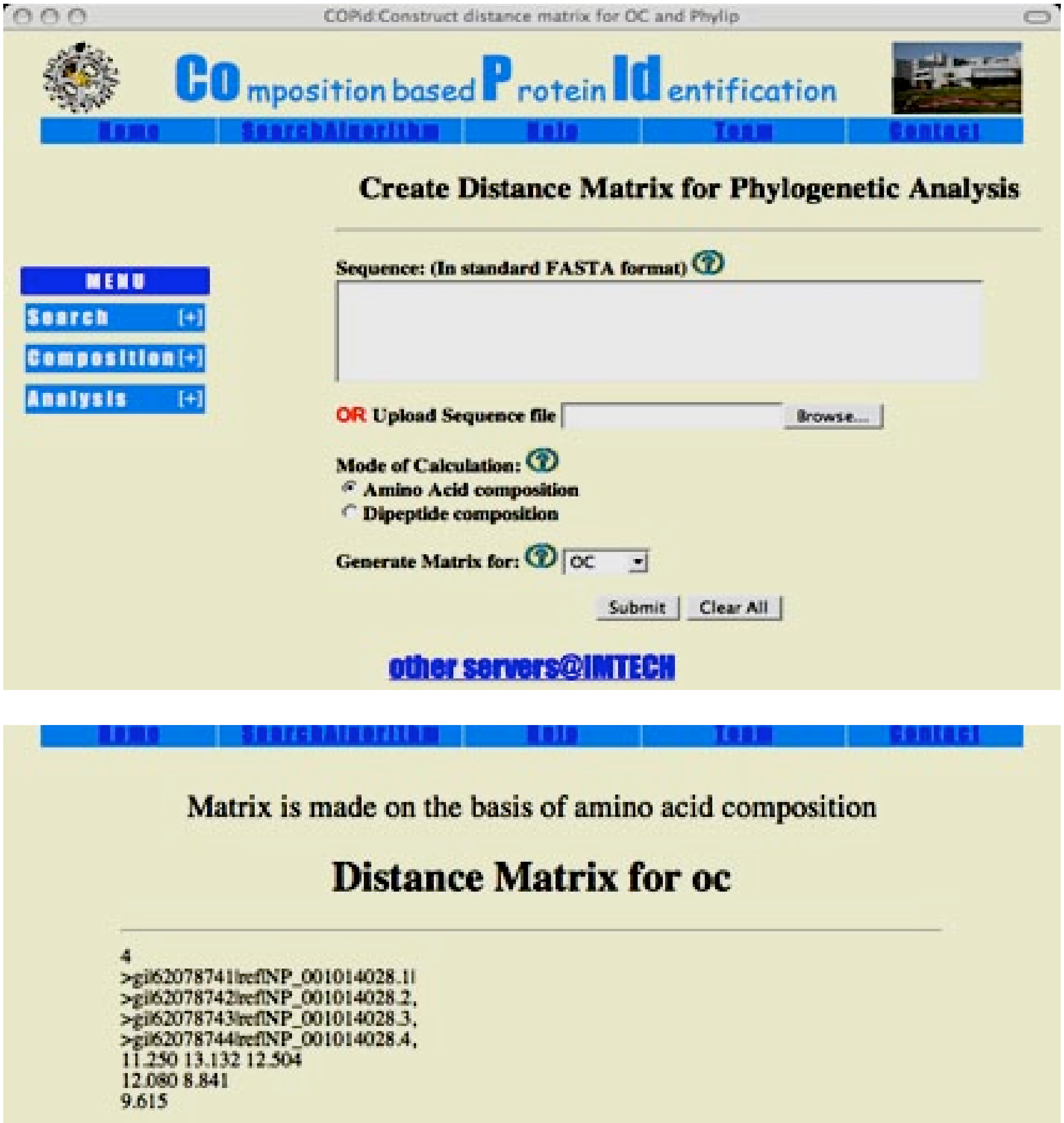
2.4.1.Click ‘phylogenetic tree’ under analysis heading on home page or unfold the analysis menu at left hand side of any submission form, then click phylogenetic tree.

2.4.2.Submit one or more sequences in FASTA format by either pasting in the text box or uploading the file.

2.4.3.Choose the composition on the basis of which distance matrix will be made.

2.4.4.Choose between the OC and WebPhylip,

2.4.5.Click submit to calculate composition or reset to delete data from the form.



Create Distance Matrix for Phylogenetic Analysis

Sequence: (In standard FASTA format)

OR Upload Sequence file

Mode of Calculation:
● Amino Acid composition
○ Dipeptide composition

Generate Matrix for: OC

Matrix is made on the basis of amino acid composition

Distance Matrix for oc

```
4
>gi|62078741|ref|NP_001014028.1|
>gi|62078742|ref|NP_001014028.2,
>gi|62078743|ref|NP_001014028.3,
>gi|62078744|ref|NP_001014028.4,
11.290 13.132 12.504
12.080 8.841
9.615
```

Figure 4: Submission form for construction of distance matrix by COPid (upper panel) and distance matrix (lower panel). The distance matrix calculated by COPid can be directly submitted to OC and WebPhylip for making of phylogenetic tree.

2.5.Comparision between two group of sequences (Figure 5a-c):

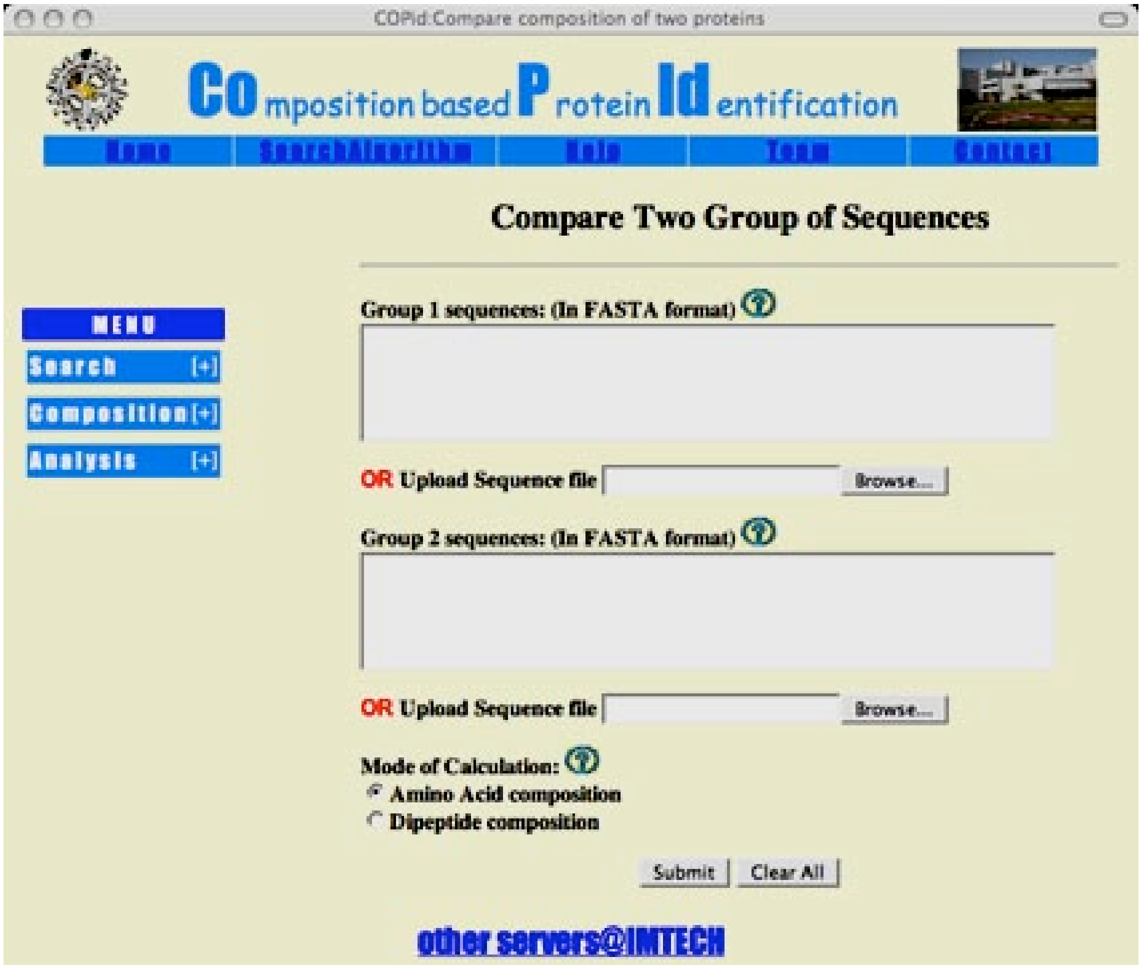
2.5.1.Click ‘compare two group of sequences’ under analysis heading on home page or unfold the analysis menu at left hand side of any submission form, then click ‘compare two group of sequences’.

2.5.2.Submit first group of sequences in FASTA format by either pasting in the text box or uploading the file.

2.5.3.Submit second group of sequences in FASTA format by either pasting in the text box or uploading the file.

2.5.4.Select between amino acid and dipeptide composition.

2.5.5.Click submit to calculate composition or reset to clear all data from the form.



Compare Two Group of Sequences

Group 1 sequences: (In FASTA format)

OR Upload Sequence file

Group 2 sequences: (In FASTA format)

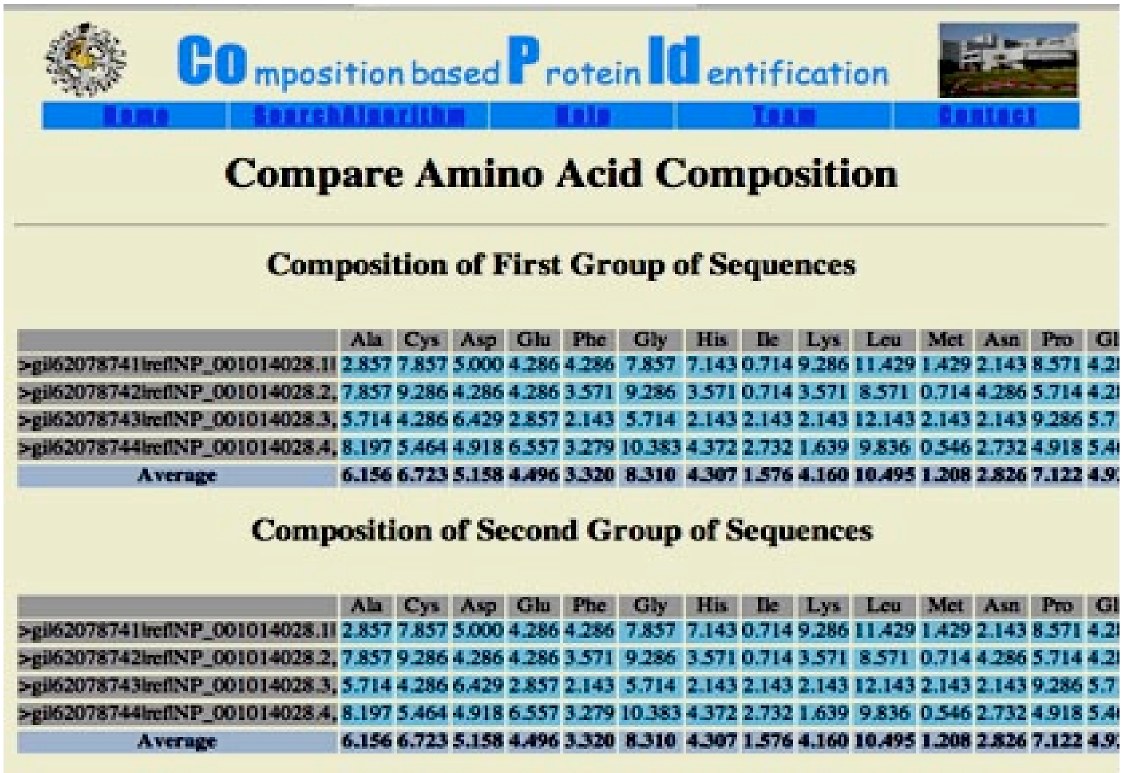
OR Upload Sequence file

Mode of Calculation:

● Amino Acid composition

○ Dipeptide composition

Figure 5a: Submission form to compare two group of sequences.



CO mposition based P rotein Id entification

Compare Amino Acid Composition

Composition of First Group of Sequences

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gl
>gi|62078741|ref|NP_001014028.1|	2.857	7.857	5.000	4.286	4.286	7.857	7.143	0.714	9.286	11.429	1.429	2.143	8.571	4.2
>gi|62078742|ref|NP_001014028.2,	7.857	9.286	4.286	4.286	3.571	9.286	3.571	0.714	3.571	8.571	0.714	4.286	5.714	4.2
>gi|62078743|ref|NP_001014028.3,	5.714	4.286	6.429	2.857	2.143	5.714	2.143	2.143	2.143	12.143	2.143	2.143	9.286	5.7
>gi|62078744|ref|NP_001014028.4,	8.197	5.464	4.918	6.557	3.279	10.383	4.372	2.732	1.639	9.836	0.546	2.732	4.918	5.4
Average	**6.156**	**6.723**	**5.158**	**4.496**	**3.320**	**8.310**	**4.307**	**1.576**	**4.160**	**10.495**	**1.208**	**2.826**	**7.122**	**4.9**

Composition of Second Group of Sequences

	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gl
>gi|62078741|ref|NP_001014028.1|	2.857	7.857	5.000	4.286	4.286	7.857	7.143	0.714	9.286	11.429	1.429	2.143	8.571	4.2
>gi|62078742|ref|NP_001014028.2,	7.857	9.286	4.286	4.286	3.571	9.286	3.571	0.714	3.571	8.571	0.714	4.286	5.714	4.2
>gi|62078743|ref|NP_001014028.3,	5.714	4.286	6.429	2.857	2.143	5.714	2.143	2.143	2.143	12.143	2.143	2.143	9.286	5.7
>gi|62078744|ref|NP_001014028.4,	8.197	5.464	4.918	6.557	3.279	10.383	4.372	2.732	1.639	9.836	0.546	2.732	4.918	5.4
Average	**6.156**	**6.723**	**5.158**	**4.496**	**3.320**	**8.310**	**4.307**	**1.576**	**4.160**	**10.495**	**1.208**	**2.826**	**7.122**	**4.9**

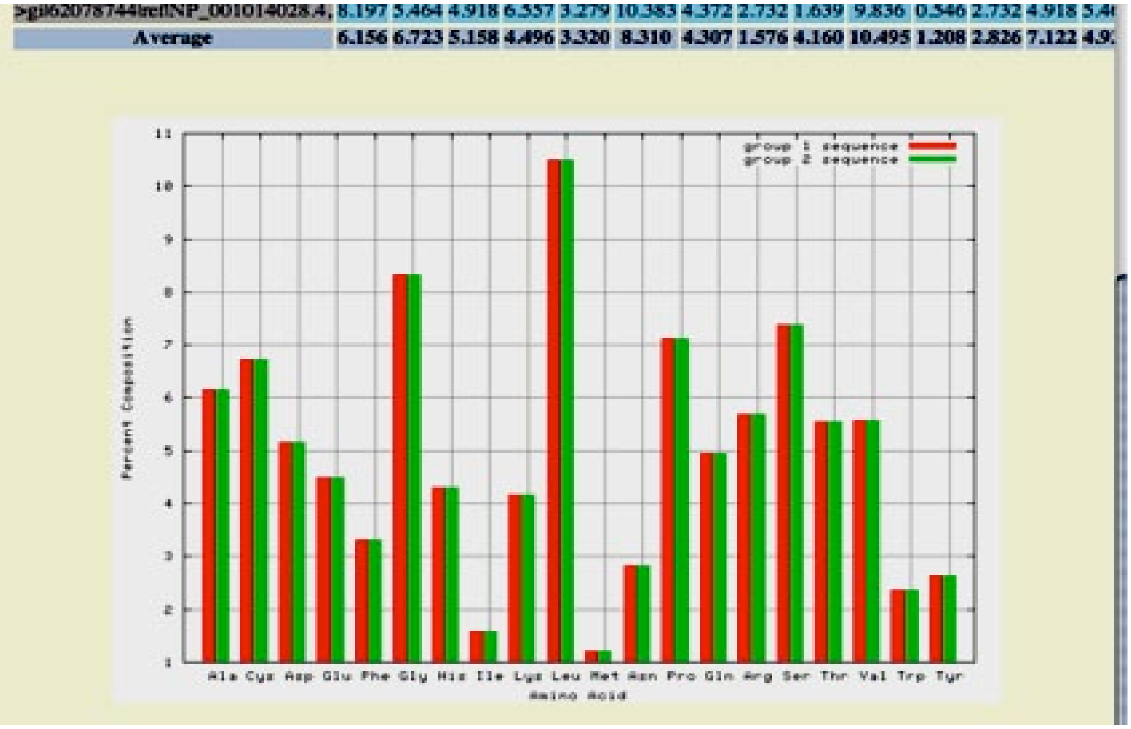
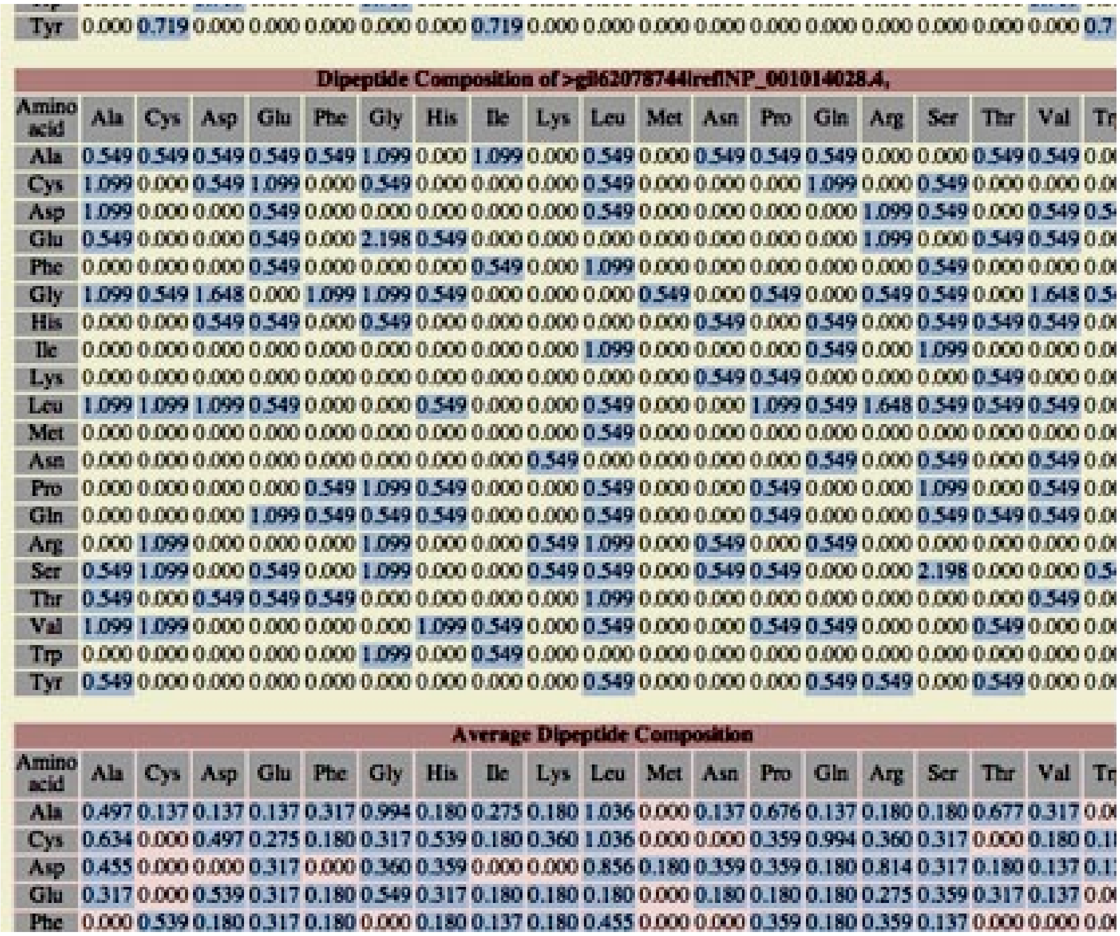

>gi|62078744|ref|NP_001014028.4,	8.197	5.464	4.918	6.557	3.279	10.383	4.372	2.732	1.639	9.836	0.546	2.732	4.918	5.4
Average	**6.156**	**6.723**	**5.158**	**4.496**	**3.320**	**8.310**	**4.307**	**1.576**	**4.160**	**10.495**	**1.208**	**2.826**	**7.122**	**4.9**

Figure 5b: Comparision of two group of sequences. Individual as well as average composition is also displayed (upper panel). Beside this average composition is also displayed in form of a bar chart.



Tyr	0.000	0.719	0.000	0.000	0.000	0.000	0.000	0.719	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.7

Dipeptide Composition of >gi|62078744|ref|NP_001014028.4,

Amino acid	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Tr
Ala	0.549	0.549	0.549	0.549	0.549	1.099	0.000	1.099	0.000	0.549	0.000	0.549	0.549	0.549	0.000	0.000	0.549	0.549	0.0
Cys	1.099	0.000	0.549	1.099	0.000	0.549	0.000	0.000	0.000	0.549	0.000	0.000	0.000	1.099	0.000	0.549	0.000	0.000	0.0
Asp	1.099	0.000	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.000	1.099	0.549	0.000	0.549	0.5
Glu	0.549	0.000	0.000	0.549	0.000	2.198	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.099	0.000	0.549	0.549	0.0
Phe	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.549	0.000	1.099	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.0
Gly	1.099	0.549	1.648	0.000	1.099	1.099	0.549	0.000	0.000	0.000	0.549	0.000	0.549	0.000	0.549	0.549	0.000	1.648	0.5
His	0.000	0.000	0.549	0.549	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.549	0.000	0.549	0.549	0.549	0.0
Ile	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.099	0.000	0.000	0.000	0.549	0.000	1.099	0.000	0.000	0.0
Lys	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.549	0.000	0.000	0.000	0.549	0.000	0.0
Leu	1.099	1.099	1.099	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.000	0.000	1.099	0.549	1.648	0.549	0.549	0.549	0.0
Met	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
Asn	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.000	0.549	0.000	0.549	0.000	0.549	0.0
Pro	0.000	0.000	0.000	0.000	0.549	1.099	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.000	0.000	1.099	0.000	0.549	0.0
Gln	0.000	0.000	0.000	1.099	0.549	0.549	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.000	0.000	0.549	0.549	0.549	0.0
Arg	0.000	1.099	0.000	0.000	0.000	1.099	0.000	0.000	0.549	1.099	0.000	0.549	0.000	0.549	0.000	0.000	0.000	0.000	0.0
Ser	0.549	1.099	0.000	0.549	0.000	1.099	0.000	0.000	0.549	0.549	0.000	0.549	0.549	0.000	0.000	2.198	0.000	0.000	0.5
Thr	0.549	0.000	0.549	0.549	0.549	0.000	0.000	0.000	0.000	1.099	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.0
Val	1.099	1.099	0.000	0.000	0.000	0.000	1.099	0.549	0.000	0.549	0.000	0.000	0.549	0.549	0.000	0.000	0.549	0.000	0.0
Trp	0.000	0.000	0.000	0.000	0.000	1.099	0.000	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.0
Tyr	0.549	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.549	0.549	0.000	0.549	0.000	0.0

Average Dipeptide Composition

Amino acid	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Tr
Ala	0.497	0.137	0.137	0.137	0.317	0.994	0.180	0.275	0.180	1.036	0.000	0.137	0.676	0.137	0.180	0.180	0.677	0.317	0.0
Cys	0.634	0.000	0.497	0.275	0.180	0.317	0.539	0.180	0.360	1.036	0.000	0.000	0.359	0.994	0.360	0.317	0.000	0.180	0.1
Asp	0.455	0.000	0.000	0.317	0.000	0.360	0.359	0.000	0.000	0.856	0.180	0.359	0.359	0.180	0.814	0.317	0.180	0.137	0.1
Glu	0.317	0.000	0.539	0.317	0.180	0.549	0.317	0.180	0.180	0.180	0.000	0.180	0.180	0.180	0.275	0.359	0.317	0.137	0.0
Phe	0.000	0.539	0.180	0.317	0.180	0.000	0.180	0.137	0.180	0.455	0.000	0.000	0.359	0.180	0.359	0.137	0.000	0.000	0.0

Figure 5c: Dipeptide composition of two group of sequences. As similar to amino acid composition, first composition of individual protein will be displayed followed by their average. But unlike amino acid composition, no graphical comparision will be shown.

2.6.Generation of amino acid or dipeptide composition for training of SVM, SNNS and timbl:

2.6.1.Click ‘pattern generation’ under analysis heading on home page or unfold the analysis menu at left hand side of any submission form, then click ‘creation of patterns for different softwares.

2.6.2.Enter name of job (optional).

2.6.3.Enter e-mail id on which you want to get information regarding completion of job (optional).

2.6.4.Submit sequences of positive dataset in FASTA format by either pasting in the text box or uploading the file.

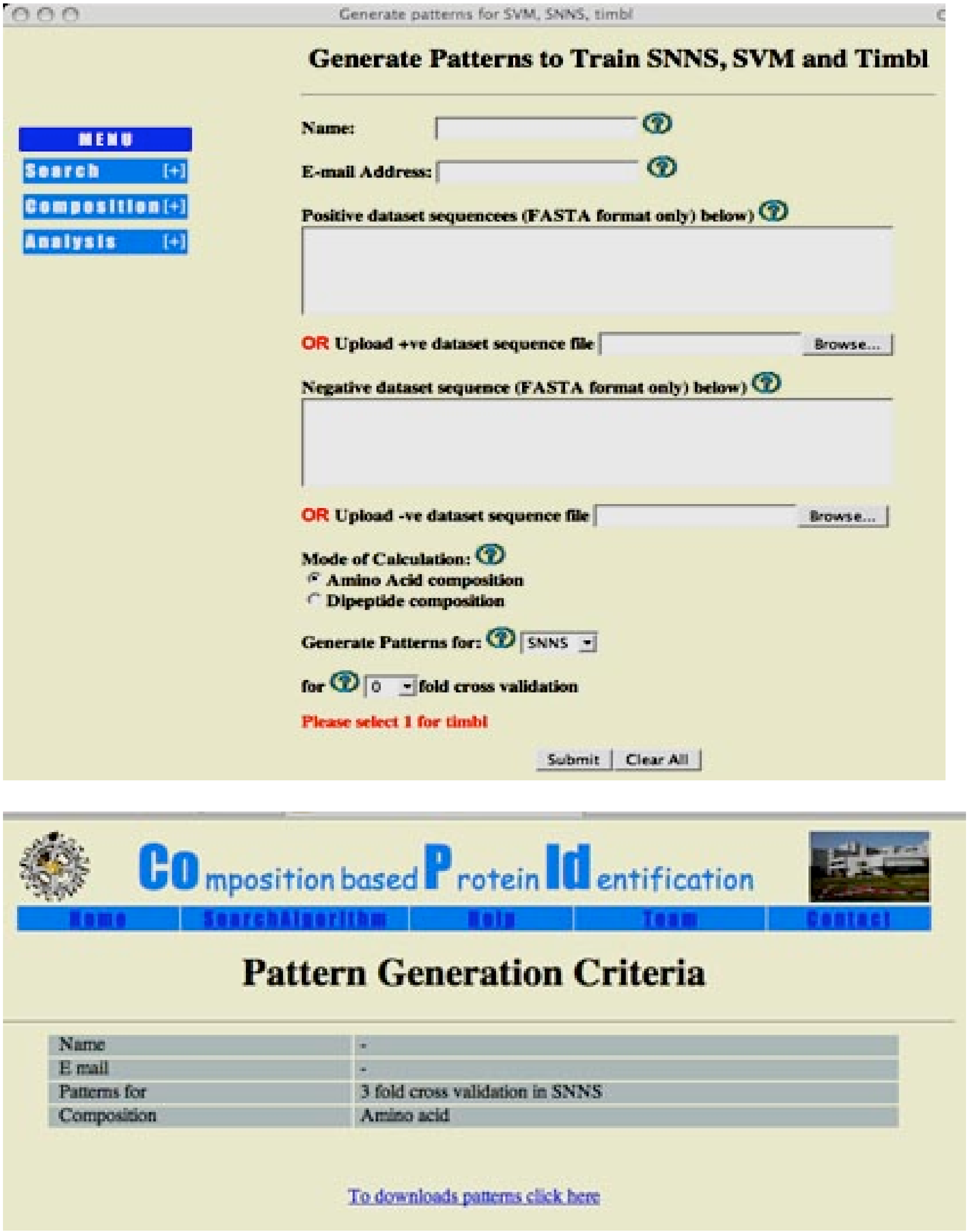
2.6.5.Submit sequences of negative dataset in FASTA format by either pasting in the text box or uploading the file.

2.6.6.Select the composition which pattern is required.

2.6.7.Select one among SNNS, SVM and Timbl.

2.6.8.Specify the number of cross fold, according to which pattern will be made.

2.6.9.Click submit to start searching or reset to clear all data in the form.



Generate Patterns to Train SNNS, SVM and Timbl

Name:

E-mail Address:

Positive dataset sequencees (FASTA format only) below)

OR Upload +ve dataset sequence file Browse...

Negative dataset sequence (FASTA format only) below)

OR Upload -ve dataset sequence file Browse...

Mode of Calculation:
● Amino Acid composition
○ Dipeptide composition

Generate Patterns for: SNNS

for 0 fold cross validation

Please select 1 for timbl

Submit | Clear All

Pattern Generation Criteria

Name	-
E mail	-
Patterns for	3 fold cross validation in SNNS
Composition	Amino acid

To downloads patterns click here

Figure 6: Submission form for generation of patterns for SVM, SNNS and Timbl (upper panel). The patterns are made available in for of tar file, which can be downloaded from COPid server.

3. Notes:

To avoid the misuse of the site the services are available for the registered users only. Users who are interested to use these servers are required to register themselves at www.imtech.res.in/errors/noauth.html. They need to fill up a registration form if they agree to the terms and conditions stated in the form. The user name and password is then sent by e-mail to the users.