MAgPI: A Memetic Algorithm Based Approach in Protein Inference Problem

Abstract—**The shotgun proteomics strategy, based on digesting proteins into peptides and sequencing them using tandem mass spectrometry along with automated database searching has become widely accepted choice of method for identifying proteins in most large-scale studies. However, the peptide centric nature of shotgun proteomics complicates the analysis and biological interpretation of the data, especially in the case of higher eukaryote organisms. The same peptide sequence can be present in multiple different proteins or protein isoforms. Such shared peptides therefore can lead to ambiguities in determining the identities of sample proteins. The protein inference problem represents a major challenge in proteomics research. In this paper, we proposed memetic algorithmic approach i.e. MAgPI to infer proteins from a given peptide sequence identified from tandem mass spectrometry in shotgun proteomics. We mapped the protein inference problem as an evolutionary search problem, and our proposed method accommodate evolutionary search with Baldwinian effect to solve the problem. In the sear strategy selected individuals are allowed to learn from surrounding and are allowed to propagate the improvement to the subsequent generations. We used Sigma49 dataset to test our method and found the proposed algorithm better than the existing approaches in literature.**

Keyword - Bioinformatics, Memetic Algorithm, Protein Inference, Evolutionary Algorithm

I. Introduction

Protein identification plays a vital role in investigating several bio-physical incidents in the field of medical science, pharmacy, molecular biology, bio-informatics and computational data analysis industry. Protein Inference is an intermediate step of the protein identification problem. It has gained immense attention in the study of proteins can provide crucial information that is not trivially available from genomic sequence. Protein inference is concerned with the identification of proteins present in a given cell or tissue which is an explicit goal of proteomics research. One of the most promising developments to come from the study of proteins has been the identification of potential new drugs for the treatment of disease. This relies on genome and proteome information to identify proteins associated with a disease, which computer software can then use as targets for new drugs. The better the accuracy of protein inferencing the greater is its impact on proteomics research. Various medical and data analysis tools have been devised to perform this task, but yet the accuracy and perfectness of the inference is a major concern and to be solved issue up to now.

Given a biological cell or tissue, the problem of determining which proteins are present in the sample is referred to as the protein inference problem. Now, the protein identification process may be divided into three sequential steps: *peptide identification*, *protein inference* and *result evaluation*. Through a sequence of experimental and computational procedure [14], [15], a given biological sample is analysed. Finally, tandem mass spectrometry (MS/MS) [9] is typically used to separate, fragment and finally identify [16][17] peptides from the sample. After the peptides have been identified, it is necessary to validate the identification process so that further steps do not suffer from noisy inputs. If there were a one to one mapping from the proteins to peptides or vice-versa, the solution would be trivial, but the actual situation is not that simple. Even if the identified set of peptides are reliable, it does not ensure that a reliable list of proteins can be assembled from these peptides. Two terminologies that are commonly encountered in protein inference problem are: one-hit wonders and degenerate peptides [13]. A protein is said to be a one-hit wonder if it generates only a single peptide when digested. It would seem that a one-hit wonder is easy to infer, but the presence of a false positive is highly likely and thus, the claim of a particular peptide being present is not absolute and protein inference faces the problem of noisy data. On the other hand, if a peptide is common between multiple proteins then that peptide is called a degenerate peptide. Thus, in the presence of a degenerate peptide, it is always difficult to deduce with protein or protein family actually generated the peptide. As a result, the protein inference problem [18], often has multiple solutions and can be computationally intractable. Three types of model are used in the protein inference domain: statistical model (ensures protein inferencing with high accuracy), parsimonious model (assumes only a small subset of proteins should be sufficient to explain all identified peptides), and optimistic model (returns all the protein or protein families that has some potential).

The main motivation behind our approach is to reach globally optimal solution avoiding the local optima. In steady state GA approach, there is a high chance of solutions being stuck in local optima. Unlike steady state GA, MAgPI does not neglect the less fit individuals completely, rather it tries to favour them by letting them learn from their surroundings [Section III-E]. It gives some chance to the less fit individuals to reproduce [Section III-C] and even to survive [Section III-F].

The contributions of this paper are as follows:

- Our approach is based on Memetic algorithm. We explored more utilization of evolutionary computation in this research.
- We designed our approach to maintain a good balance

between the exploration and exploitation throughout the lifetime of this evolutionary search. We utilized different technique in maintaining diversity of solutions (diversity maintenance by selection scheme) from the traditional distance based diversity maintenance.
- Our approach provides the flexibility to implement the protein inference process as parsimonious model or optimistic model or somewhere between the two, based on some tuning parameter as discussed in III.
- Our approach is more statistically robust as we infer protein for various random test peptide-set (selected based on the probability of detectability) with high precision and recall.

II. Related Works

Over the years, various methods have been tried and tested to conduct the protein inference process. Nesvizhskii and colleagues first addressed this challenge using a probabilistic model [20], but different problem formulations and new solutions have been proposed as well [19], [21], [22]. A combinatorial approach to the protein inference problem that incorporates the concept of peptide detectability, i.e. the probability of a peptide to be detected (identified) in a standard proteomics experiment, with the goal of finding the set of proteins with the minimal number of missed peptides is discussed in [21]. In the other combinatorial formulation [22], the parsimony condition was chosen without theoretical justification, rather for convenience reasons only. Furthermore, parsimonious formulations often lead to the minimum cover set problem, which is NP-hard. Thus, greedy [21] or graph-pruning strategies [21] address the protein inference problem without performance guarantee. In another approach [23], the protein inference is addressed by proposing two novel Bayesian models that take as input a set of identified peptides from any peptide search engine, and attempt to find a most likely set of proteins from which those identified peptides originated. The basic model assumes that all identified peptides are correct, whereas the advanced model also accepts the probability of each peptide to be present in the sample. Few classes of cooperative meta-heuristics like the Island model, Spatially em-bedded models and genetic programming have been used in protein identification. Popitam [24], exploits the concept of spectrum graph in order to extract tags (amino acid sequences) from the MS/MS spectrum. The graph represents all possible complete sequences and sub-sequences that can possibly be built from the spectrum . Vertices are built from the peaks and represent therefore masses of fragments, while edges represent amino acid masses. The sequences can be constructed by moving from one vertex to another by following existing edges. [26], [27], [28], [29], [30] typically try to extract tags (or complete sequences) from the graph and use them to identify the most similar sequence from the database using sequence alignment algorithms [25]. The evolutionary identification approach [31] tries to find the entire sequence of a protein, even in the case of variants or unknown proteins. To accomplish that, different peptides that composes a given protein must be identified. First, their mass have to be found with a MS spectrum and secondly, from their mass, their sequence can be found with MS/MS spectra.

III. The MAgPI:Our proposed approach

Based on the motivation, we propose MAgPI(Memetic Algorithm approach for Protein Inference). We applied the Memetic Algorithm with diversity maintenance mechanism in this approach. The summary of the procedure is presented in Algorithm 1. The details are presented in the later subsections.

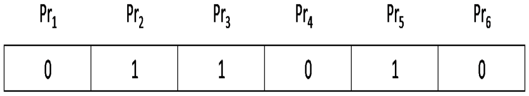
Algorithm 1 Memetic Algorithm Based Approach in Protein Inference(MAgPI)

```
Initialize population $P(0)$ with $\lambda$ solutions
$k \leftarrow 0$
while $k < G_{max}$ do
    $Q(k) \leftarrow \emptyset$
    while $|Q(k)| < \mu$ do
        $a_1 \leftarrow RWSS(P(k))$
        $a_2 \leftarrow FUSS(P(k))$
        $[o_1, o_2] \leftarrow CROSSOVER(a_1, a_2)$
        $MUTATE(o_1)$ with probability $\vartheta$
        $MUTATE(o_2)$ with probability $\vartheta$
        $Q(k) \leftarrow Q(k) \cup \{o_1, o_2\}$
    end while
    $L \leftarrow$ Set of candidates for Offspring Education from $P(k) \cup Q(k)$
    for $\forall \ell \in L$ do
        educate $\ell$ by offspring education procedure
    end for
    $P(k+1) \leftarrow \lambda$ survivors from $P(k) \cup Q(k)$
    $k \leftarrow k + 1$
end while
```

$RWS-$ Roulette Wheel Selection
$FUSS-$ Fitness Uniform Selection Scheme.

A. Representation of the candidate solution

Each candidate solution infers the proteins those are present in the sample. The simplest representation is to consider all the proteins and encode it such a way that represents whether a particular protein is present or not. Let us assume that each gene corresponds to whether a particular potential protein is present or not. If the protein is present, the corresponding gene has value 1 and if the protein is absent, the corresponding gene has value 0. Thus a string of 0 and 1's will form an individual where the number of 1's represents the number of proteins present in the sample and the individual 1's correspond to the presence of respective proteins. So, the representation will look something like Figure 1. In this figure, the candidate solution represents that the testing biological sample contains proteins P2, P3 and P5. Remember that the main database may contain many more proteins those are not potential with respect to the peptide sequences obtained from MS and MS/MS spectrum analysis [3] and thus can be discarded from the candidate solution representation as they can never appear. This point is further clarified in Section III-B.



Pr_1	Pr_2	Pr_3	Pr_4	Pr_5	Pr_6
0	1	1	0	1	0

Fig. 1. Representation of the candidate solution

B. Initial Population Generation

Given a protein, its peptide sequence is well known and available in many protein databases. Thus, by constructing an inverted index, the possible potential proteins can easily be identified given the set of identified peptides through shotgun proteomics. Thus, we need to consider only the potential proteins, not all of them, which significantly reduces the condidate solution size as well as computational complexity.

For the initial population generation, let use consider the mapping of protein to its peptide sequence as demonstrated below:

$$\begin{gathered} Pr_1 \rightarrow GKEUTR \\ Pr_2 \rightarrow LYARE \\ Pr_3 \rightarrow LVGARTHNB \\ \text{..........................} \\ \text{..........................} \\ Pr_n \rightarrow WHBAFGSTHJSYB \end{gathered}$$

So, if the protein is known, so is its peptide sequence. Now, after the MS and MS/MS spectra are analyzed through tandem mass spectrometry, one can determine the peptides generated from the parent proteins of the test sample. Though this peptide identification process is not perfect, it can be accomplished with reasonable accuracy. Now, if the inverted index idea is applied as discussed in the previous paragraph, one can easily determine the potential parent proteins expected to be present in the given sample. All we have to do is to take a peptide generated from the shotgun proteomics, then search the protein-peptide mapping database to find the potential parent proteins that might have generated the peptide. That is, if protein $P1$ generates a peptide sequence which contains the peptide G, then protein $P1$ is obviously a potential parent of peptide G. In this way, we find all the potential proteins in the sample, form a set of potential proteins removing any repetitions and discard the rest of the proteins as they are of no interest for this particular test case.

As the number of proteins those are present in the test sample can not be greater than the cardinality of the potential protein set, number of genes in an individual is kept equal to the cardinality of the potential protein set. After the number of potential proteins as well as the number of genes in an individual and also the individual potential proteins are identified, the next task is to assign a particular value to every gene within an individual. Gene value 1 represents the presence of the respective protein and gene value 0 represents the absence of that protein. So, we can now generate the individual of the initial population by just generating a random string of 0's and 1's. This helps to reduce the size of the candidate solution and reduce the search space. A sample initial population with 4 individuals in presented in Figure 2, where the first candidate solution inferes proteins P_4 and P_9, whereas the third individual infers proteins P_9 and P_{11}.

Pr_3	Pr_4	Pr_9	Pr_{11}	Pr_{14}
0	1	1	0	0
1	1	0	0	1
0	0	1	1	0
1	0	1	0	1

Fig. 2. A sample initial population

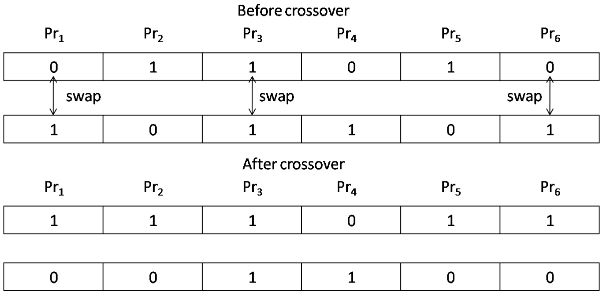
C. Parent Selection procedure

To maintain diversity and balance between exploration and exploitation two different parent selection procedure. First parent is selected based on roulette wheel selection mechanism [7]. The second parent is selected based on Fitness Uniform Selection Scheme(FUSS) [33] which select individual uniformly over fitness landscape.

D. Breeding operators

The main two breeding operators used in MAgPI are the following:

- Recombination operator
- Mutation operator



Before crossover

Pr_1	Pr_2	Pr_3	Pr_4	Pr_5	Pr_6
0	1	1	0	1	0
1	0	1	1	0	1

After crossover

Pr_1	Pr_2	Pr_3	Pr_4	Pr_5	Pr_6
1	1	1	0	1	1
0	0	1	1	0	0

Fig. 3. Uniform Crossover

Let us first consider the recombination operator. As the candidate solution is encoded as a boolean valued vector and the sequence of genes does not have any special sequence, uniform crossover [4] has been applied so that every gene has equal probability of being swapped. This swapping probability is denoted by γ which is a user defined parameter. Figure 3 demonstrates this process where uniform crossover swaps the genes corresponding to P1, P3 and P6.

A probability ϑ is used for every offspring to be mutated. For mutation, bit-flip mutation oprator is applied as the candidate solutions are represented by binary strings. That is, we randomly select a gene from an individual and with a very small probability θ, we flip the value of that gene. Figure 4 demonstrates this process where the mutation operator flips the value of P3 from 1 to 0.

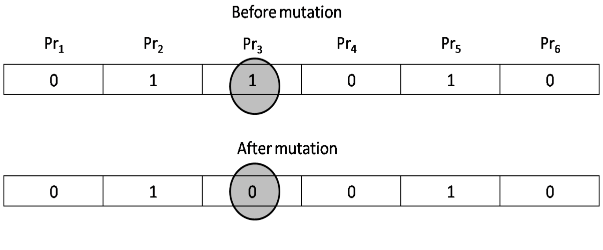


Fig. 4. Bit flip mutation operator

E. Offspring Education Procedure

To accommodate the essence of Memetic Algorithm, we have run offspring education procedure on selected individuals. $m\%$ Individuals of total population are selected for education. First $K\%$ best individuals are selected, then the rest are selected uniformly over fitness. For offspring education procedure as well as to maintain diversity we used simulated annealing method.

F. Survivor Selection procedure

The survivor selection of MAgPI is a combination of both elitist and Fitness Uniform Selection. $K\%$ of the selected individuals for next generation are selected in elitist approach, the rest are selected uniformly over fitness landscape(FUSS).

G. Fitness evaluation

While evolving the candidate solutions of protein inference problem, the fitness function should consider the following issues:

- Whether to prefer minimum number of proteins that covers the generated peptide sequence or whether to prefer the maximum number of proteins or whether to keep both the options and make it adaptive (treating it as a user defined parameter, which MAgPI ultimately does). This is one of the unique features of MAgPI which gives user the control over whether to prefer the minimum or maximum number of inferred proteins or take a mid way around.
- Percentage of peptides covered within the test peptide sequence by the inferred proteins of a candidate solution
- How much the hypothetical peptide set constructed from the inferred proteins of an individual is similar to the test peptide sequence. The more they are similar, the higher is the probability of the inferred protein set of being correct. How many redundant peptides does this hypothetical peptide set contains.
- This procedure is not deterministic due to the presence of both one-hit wonders and degenerate peptides. So, an exact solution may not be available and approximate solutions may be only possible option.

For the ease of notation, let us assume that we have the following functions:

- $N(c)$**:** Returns the number of protein present in an individual c
- $C(c)$**:** This is the coverage function which signifies number of test peptides covered by the hypothetical peptide set resulting from inferred protein set.
- $R(c)$**:** This is the redundancy function which signifies number of peptides in hypothetical peptide set resulting from inferred protein set, those are not present is test peptide set.
- $S(c)$**:** This is the shield function which returns number of test peptides not covered by the hypothetical peptide set resulting from inferred protein set.

Then the Fidelity(f) and Exposure(η) is measured using Equation 1 and 2 respectively. Fidelity signifies how trustworthy the individual is in inferring the protein according to our proposed heuristic. And Exposure signifies how much expressive an individual in expressing the test peptide set.

$$f = \frac{C(c)}{C(c) + R(c)} \tag{1}$$

$$\eta = \frac{C(c)}{C(c) + S(c)} \tag{2}$$

Now, the fitness function can be expressed in the following manner:

$$F(c) = \frac{1}{\psi_c^i * \frac{1}{f} + (1 - \psi_c^i) * \frac{1}{\eta}} * N(c)^{\epsilon} * \left(\frac{1}{N(c)}\right)^{1-\epsilon} \tag{3}$$

Where, ϵ and ψ_c^i are user defined parameters, ϵ indicates the user preference for wheather to take maximum possible number of proteins or minimum possible number of proteins as the desired output or something in between them. Note that, the value of ϵ is varied between 0 and 1. 0 is one extreme end where the user is preferring to take minimum possible number of proteins, whereas extreme end 1 means the user is preferring to take maximum possible number of proteins. ψ_c^i gives MAgPI the flexibility to prefer either high precesion or high recall or a mid way around, which the previous bayesian approches did not provide. The naive bayes approarch with multiplication of probabilities would always prefer minimum number of proteins and thus is biased towards higher precision while obtaining poor recall.

IV. Experiments

A. Dataset

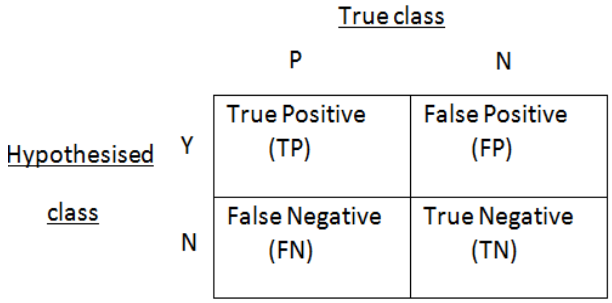
We used Sigma49 dataset collected from to test our algorithm. Sigma49 sample was prepared by mixing 49 human proteins, among which 44 proteins contain at least one peptide that can be identified by shotgun proteomics. In addition, 9 keratin proteins and 4 other proteins are categorized as the keratin contamination and bonus proteins, respectively, and are believed to be present in the sample due to contamination.Sigma49 provides 3 replicates (the same experiment run thrice) of Peptide Prophet results which contains the peptide search results based on the MS/MS spectra obtained from tandem mass spectrometry. It also contains the probabilities of the predicted peptides to be correct. These search results were used as the inputs to our approach which starts from these identified peptides and finally infers the proteins.

B. Experimental Setup

We implemented our MAgPI on a core i-3 intel microprocessor, 2 GB RAM machine using JAVA. Peptide Prophet result of sigma49 dataset was parsed to peak out identified peptides. At first probable proteins predicted by Peptide Prophet Result are given a peptide count as per Peptide Prophet result. All the peptides belonging to proteins having peptide count 2 or more are directly taken as identified peptides. For those peptides belonging to proteins with peptide count 1, predicted probability is checked and filtered with minimum probability 0.97.

C. Performance measures

The performance measures mainly used in this work are True Positive, False Positive, False Negative, precision, recall and F-measure [8]. They are briefly discussed below in reference with the confusion matrix shown in figure 5.



		True class P	True class N
Hypothesised class	Y	True Positive (TP)	False Positive (FP)
	N	False Negative (FN)	True Negative (TN)

Fig. 5. Confusion matrix

Precision: Portion of correctly inferred proteins within the test peptide set that are inferred to be present by the protein inference algorithm.

$$precision = \frac{TP}{TP + FP} \tag{4}$$

Recall: It is also called the true positive rate. It is the portion of peptides within the test peptide set that are correctly inferred by the protein inference algorithm.

$$recall = \frac{TP}{P} \tag{5}$$

F-measure: A measure that combines precision and recall both. It is their harmonic mean.

$$f\_measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{6}$$

D. Parameters

The parameters used in this research are, ψ_c^i which is used in determining fitness value from fidelity(f) and exposure(η) and is ranged $[0, 1]$. ϵ: the selection parameter for parsimony and optimism in inferring proteins also ranged $[0, 1]$, λ: the size of population, μ the size of offspring in each generation, G_{max}: maximum allowed generation, γ: uniform crossover probability, ϑ: mutation probability, θ: the probability of every bit to be flipped in mutation, m: percentage of population for offspring education, K: elitist ratio in selection both for offspring education and survivor selection. The values of different parameters used in this algorithm is noted in Table I

Parameter name	value
ψ_c^i	0.23
ϵ	0.55
λ	100
μ	100
G_{max}	100
γ	0.5
ϑ	0.7
θ	0.1
m	0.4
K	0.1

TABLE I
DIFFERENT PARAMETERS AND THEIR VALUES

E. Results

Protein inference results on the Sigma49 dataset using minimum missed peptide approach (MMP), ProteinProphet (PP) , basic Bayesian model (BB), basic Bayesian model with detectability adjustment (BBA), advanced Bayesian model using raw PeptideProphet probabilities(ABP), ABP after detectability adjustment (ABPA) advanced Bayesian model using converted probability scores (ABL), ABL after detectability adjustment (ABLA), and ABLA with estimated protein prior probabilities (ABLAP)and Meta-heuristic approach (PIssGA) and our approach (MAgPI). All results are evaluated based on the true positive (TP), false positive (FP) and false negative (FN), precision (PR), recall RC) and F-measure (F) as discussed in Section IV-C. Table II shows the summary result.
In our experimental setup we tried to infer 44 proteins of Sigma49 as the rest five proteins did not contain a single peptide that could be identified by the Peptide Prophet search. The result shows that, our approach outperforms all other approaches in terms of precision. In terms of recall it has not performed good. This is because we chose the parameters to emphasize on providing a solution that is highly precise. Number of false positive proteins in our experimental setup is less than all other approaches in literature. Which results in the highest precision of our algorithm. In this settings our this approach of protein inference gives the best result in terms of f-measure from all other existing approaches.

	MMP	PP	BB	BBA	ABP	ABPA	ABL	ABLA	ABLAP	PIssGA	MAgPI
TP	39	41.5	39	37	35	43	37	44	43	40.5	39.62
FP	6	7.5	16	6	4	22	4	9	6	8	0.3
FN	5	2.5	5	7	9	1	7	0	1	3.5	4.38
Pr	0.87	0.85	0.71	0.86	0.9	0.66	0.9	0.83	0.88	0.84	0.99
Rc	0.89	0.94	0.89	0.84	0.8	0.98	0.84	1.0	0.98	0.92	0.90
F	0.88	0.89	0.79	0.85	0.84	0.79	0.87	0.91	0.92	0.88	0.94

TABLE II
COMPARISON OF RESULTS

V. CONCLUSION

In this paper we have presented MAgPI to solve the protein inference problem and measured its performance on Sigma49 dataset. The experimental results show that, our algorithm outperforms all of the existing protein inference methods. Given a peptide sequence, our algorithm is drastically faster in terms of identifying proteins than the existing approaches. MAgPI utilizes a new dimension of evolutionary computation in protein inference problem. While proceeding in this approach, we were also concerned about maintenance of diversity to avoid premature convergence in local optima. Instead of using traditional distance based diversity maintenance we utilized selection operator based diversity maintenance(FUSS) method. The results obtained so far are promising and further investigation with other variants of evolutionary computations may yield better results. Another direction can be to formulate the problem as a multi-objective optimization.

ACKNOWLEDGEMENT

This research is a continuation of “PIssGA: An ultra fast meta-heuristic approach to solve protein inference problem”, and content of this paper is slightly adopted from that paper.

REFERENCES

[1] Fundamentals of scatter search and path relinking. F Glover, M Laguna-Control and cybernetics, 2000 - leeds-faculty.colorado.edu

[2] Evolution strategies A comprehensive introduction [PDF] from bham.ac.uk HG Beyer - Natural computing, 2002 - Springer

[3] Alexey I. Nesvizhskii[1], Ruedi Aebersold[2]. Interpretation of Shotgun Proteomics Data: The Protein Inference Problem. [1]Institute for System s Biology,1441 N 34th Street, Seattle WA 98103 USA. [2]ETHZ Federal Institute of Technology, CH-8092 Zurich, Switzerland

[4] Uniform Crossover in Genetic Algorithms by: Gilbert Syswerda edited by: David J. Schaffer In Proceedings of the Third International Conference on Genetic Algorithms (1989), pp. 2-9

[5] Genetic Algorithms, Tournament Selection, and the E ects of Noise [PDF] from psu.edu BL Miller- Urbana, 1995 - Citeseer

[6] Error correcting codes in quantum theory [PDF] from princeton.edu AM Steane - Physical Review Letters, 1996 - APS

[7] D. E. Goldberg and K. Deb. A comparative analysis of selection schemes used in genetic algorithms. InFoundations of Genetic Algorithms, pages 69-73. Morgan Kaufmann, 1991.

[8] Tom Fawcett . An introduction to ROC analysis

[9] Alexey I. Nesvizhskii[1], Ruedi Aebersold[2]. Interpretation of Shotgun Proteomics Data: The Protein Inference Problem. [1]Institute for System s Biology,1441 N 34th Street, Seattle WA 98103 USA. [2]ETHZ Federal Institute of Technology, CH-8092 Zurich, Switzerland

[10] MA Le Gros, G McDermott. X-ray tomography of whole cells. Current opinion in structural,2005 - Elsevier.

[11] Nicholas T. Hartman, Francesca Sicilia, Kathryn S. Lilley, and Paul Dupree. Proteomic Complex Detection Using Sedimentation. Department of Biochemistry, University of Cambridge, Building O, Downing Site, Cambridge CB2 1QW, United Kingdom.

[12] MR Emmert-Buck, RF Bonner, PD Smith. Laser capture microdissection. Science, 1996 - sciencemag.org.

[13] T Huang, J Wang, W Yu Protein inference: a review Briefings in Bioinformatics, 2012 - Oxford Univ Press

[14] Aebersold, R., Mann, M. Mass spectrometry-based proteomics. Nature 422, 198207 (2003)

[15] Kislinger, T., Emili, A. Multidimensional protein identification technology: current status and future prospects. Expert Rev. Proteomics 2(1), 2739 (2005)

[16] Marcotte, E.M. How do shotgun proteomics algorithms identify proteins?. Nat. Biotechnol. 25(7), 755757 (2007)

[17] Nesvizhskii, A.I. Protein identification by tandem mass spectrometry and sequence database searching. Methods Mol Biol 367, 87119 (2007)

[18] Nesvizhskii, A.I., Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4(10), 14191440 (2005)

[19] Nesvizhskii, A.I., Aebersold, R.: Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4(10), 14191440 (2005)

[20] Nesvizhskii, A.I.,Keller, A., Kolker, E., Aebersold, R.: A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75(17), 46464658 (2003)

[21] Alves, P., Arnold, R.J., Novotny, M.V., Radivojac, P., Reilly, J.P.,Tang, H.: Advancement in protein inference from shotgun proteomics using peptide detectability. In: PSB 2007: Pacific Symposium on Biocomputing, pp. 409420. World Scientific, Singapore (2007)

[22] Zhang, B., Chambers, M.C., Tabb, D.L.: Proteomic Parsimony through Bipartite Graph Analysis Improves Accuracy and Transparency. J Proteome Res. 6(9), 35493557 (2007)

[23] Yong Fuga Li, Randy J. Arnold, Yixue Li, Predrag Radivojac, Quanhu Sheng, and Haixu Tang1: A Bayesian Approach to Protein Inference Problem in Shotgun Proteomics. RECOMB 2008, LNBI 4955, pp. 167180, 2008.

[24] P. Hernandez, R. Gras, J. Frey, and R.D. Appel. Popitam: Towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics, 3(6):870-879, 2003.

[25] A.J. Mackey, T.A. Haystead, and W.R. Pearson. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. Mol. Cell Proteomics, 1:139-147, 2002.

[26] V. Dancik, T. Addona, K. Clauser, J. Vath, and P.A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol., 6:327-342, 1999.

[27] T. Chen, M.Y. Kao, M. Tepel, J. Rush, and G.M. Church. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. J. Comput. Biol., 8(3):325-”337, 2001.

[28] M. Mann and M.Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem, 66:4390”4399, 1994.

[29] A. Schlosser and W.D. Lehmann. Patchwork peptide sequencing: extraction of sequence information from accurate mass data of peptide tandem mass spectra recorded at high resolution. Proteomics, 2:524-533, 2002.

[30] J.A. Taylor and R.S. Johnson. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom, 11:1067-1075, 1997.

[31] Jean-Charles Boisson, Laetitia Jourdan and El-Ghazali Talbi, Christian Rolando. A Prelimainary Work on Evolutionary Identification of Protein Variants and New Proteins on Grids. In Proceedings of the 20th International Conference on Advanced Information Networking and Applications 2006.

[32] Shubhra Kanti Karmaker Santu, S. Rahman, Saikat Chakraborty , and M. Sohel Rahman, PIssGA: An ultra fast meta-heuristic approach to solve protein inference problem. In Proceedings of 16th International Conference on Computer and Information Technology (ICCIT), 2013

[33] Marcus Hutter, Fitness Uniform Selection to Preserve Genetic Diversity, on Technical Report IDSIA-01-01, 17 January 2001, IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland