

Mass Spectrometry and Computational Proteomics

Vineet Bafna
Computer Science & Engineering,
Univ. California, San Diego, AP&M3832
9500 Gilman Drive,
La Jolla, CA 92093-0114, USA.
Email: vbafna@cs.ucsd.edu.
858-822-4978(W)

Knut Reinert
Algorithmic Bioinformatics,
Institut for Computer Science,
Free University Berlin, 14195
Berlin, Germany
Email: reinert@mi.fu-berlin.de
49-30-838-75222(W)

July 13, 2004

Abstract

Mass Spectrometry is the tool of choice for Proteomics, with applications to peptide sequencing, protein structure prediction, protein-protein interactions, and many others. Continued improvements in instrumentation and computational technologies will only help accelerate this trend. A short overview of algorithms for interpreting mass spectrometry (MS) data is provided. This overview is not intended as an introduction to the technology itself or to proteomics. Instead, an abstract overview of MS data is presented in order to describe key algorithmic ideas required for its interpretation. Three proteomic applications are considered: *protein identification*, *protein interactions*, and *differential analysis of protein expression*.

Keywords: *de novo* sequencing, Dynamic Programming, Electrospray, MALDI, Mass Spectrometry, Tandem Mass Spectrometry, Proteomics.

Introduction

Computation occupies a central role in the interpretation of high throughput biological data, and Proteomics is no exception. It can also be argued that *mass spectrometry* is now the tool of choice for proteomics, with applications to peptide sequencing, protein structure prediction, protein-protein interactions, (relative) protein expression, and many others. Continued improvements in instrumentation and computational technologies will only help accelerate this trend. Indeed, the chemistry Nobel prize in 2002 was awarded for the development of protein ionization methods in mass spectrometry. While rewarding individual accomplishments, the awards were also a recognition of the immense potential of this field.

In this overview, we summarize algorithms for interpreting mass spectrometry (MS) data. This short overview is not intended an introduction to the technology itself or to proteomics (see for example [27, 31]). Instead, we will provide an abstract overview of MS data in order to describe key algorithmic ideas required for its interpretation. Due to space limitations, we will concentrate on three applications: *protein identification*, *protein interactions*, and *differential analysis of protein expression*.

MS technology

The mass spectrometer is a device that measures the mass (actually, the mass to charge ratio) of an ionized molecule. Its key components are a source for sample introduction and ionization (typically MALDI (Matrix Assisted Laser Desorption Ionization) or ESI (ElectroSpray Ionization)), and a *mass analyzer* for measuring the mass (e.g. Ion Trap, TOF (time of flight), Quadrupole, and Fourier transform (FT-MS)). The conceptual question one might ask is the following: how can a device that essentially measures mass be used in diverse applications like protein sequencing, expression and structure?

To answer this, one must note that while these devices are conceptually simple, they now offer extremely high mass accuracy and resolving power. Various protocols can be used to obtain a set of characteristic masses that would be diagnostic for a protein. In *Peptide Mass Fingerprinting* (PMF), the protein is enzymatically digested, and the peptide masses are recorded using a single mass spectrometric measurement (MS). Every protein will have a characteristic set of digested peptides, and correspondingly, peptide masses. For more complex mixtures this procedure is not sufficient and instead multiple stages of mass spectrometry are applied (Tandem MS (MS/MS) or MS^n).

In tandem mass spectrometry, the peptides (from an enzymatically digested protein mixture) are ionized with one or more units of charge, as in single stage MS, and a specific peptide is chosen for fragmentation by *collision-induced dissociation* (CID). Fragments retaining the ionizing charge after CID have their mass-charge ratio measured in a second stage of mass spectrometry. Since peptides typically break a peptide-bond when they fragment by CID, the resulting spectrum contains information about the constituent amino-acids of the peptide.

The fragmentation of the peptide in CID is a stochastic process governed by the physiochemical properties of the peptide and the energy of collision. The charged fragment can be inferred by the position of the broken bond and the side retaining the charge. In figure 1(b), the N-terminal a_1, b_1, c_1 fragment ions, and the C-terminal x_{n-1}, y_{n-1} , and z_{n-1} fragment ions are shown. While a, b, y represent the commonly occurring fragments, a high energy collision often results in other fragments, including *internal* fragments formed by breakage at two points, fragments formed by breaks in side-chains, and neutral molecule losses from fragments, including H_2O , and NH_3 . One or more of these fragments retain the charge unit(s), and their mass-charge ratio is registered. Figure 1(c) shows the single charge being retained by y_{n-1} . In a single experiment, many charged fragments are formed by CID of multiple copies of the same peptide. The aggregate of the mass-charge ratios detected is called the *MS/MS spectrum*. A cartoon MS/MS spectrum for the peptide SGFLEEDK is shown in Figure 2. It helps illustrate how the MS/MS spectrum can be used to determine the sequence of amino-acids of a peptide. Note that the difference in mass-charge ratio of the adjacent singly-charged y -ions, y_5 , and y_6 is exactly the mass of the residue F . The algorithmic challenge is to use the MS/MS data in the presence of noise, incomplete fragmentation, and mixed prefix, suffix, and internal ions, to provide unambiguous identification.

Protein Identification

MS/MS might not be necessary if a relatively simple mixture of proteins is being analyzed. Instead, PMF can be used with every protein containing a characteristic set of peptide masses. Two problems complicate this simple picture. First, only a small fraction of the peptides are typically ionized and detected by MS. Second, the observed peptides are sometimes from multiple (2 – 10) dominant proteins in the mixture. These issues are partially resolved using a Bayesian probabilistic model that assigns a likelihood of a protein sequence given a spectrum of peptide masses [24]. The protein with the highest likelihood is then reported. In order to deal with multiple proteins, the process is repeated after 'removing' the peaks that matched the first protein. For

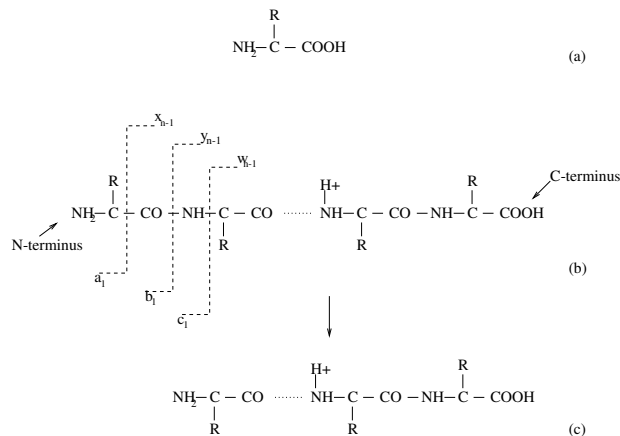


Figure 1: (a) The structure of an amino-acid. (b) An ionized peptide. (c) y_{n-1}^+ ion

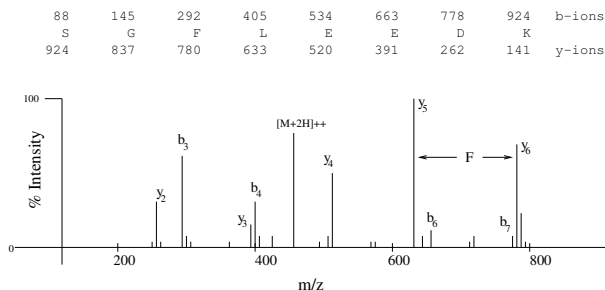


Figure 2: MS/MS spectrum for peptide SGFLEEDK.

many proteomic applications with complex mixtures of proteins, single stage MS is gradually being supplanted by tandem, and higher order mass spectrometry as the more reliable method for protein sequencing.

The core of most of the software programs for analyzing tandem MS data contain an implementation of the following three modules: *interpretation*, *filtering* and *scoring*. In the interpretation module, the input is an MS/MS spectrum, and the output is an *interpreted-spectrum* containing meta-information that can be reliably inferred from the spectrum. It may include parent peptide mass, partial or complete sequence tags, and combinations of sequence tags and molecular masses. The filtering module takes the interpreted-spectrum and a peptide sequence database as input, and filters out most of the peptides, leaving a small list of *candidate-peptides* that might have generated the MS/MS spectrum. Finally, the scoring module takes the list of candidate-peptides MS/MS spectrum as input, and returns a ranking of the *candidate-peptides* along with a score and possibly a p -value (probability that the score was achieved by random chance). If significant, the highest scoring peptide is the correct interpretation. As the peptide sequences have relatively low redundancy, the identification of the sequence of one or two peptides is usually sufficient to identify the protein. Certainly, *scoring* is the key module and is under active research [1, 13, 14, 28]. However, most existing toolkits use some aspects of all three modules. With an increase in the size of databases, and growing interest in post-translational modifications, algorithms for interpretation and filtering are coming to the fore.

De novo Interpretation

We take a general view by defining interpretation to be any sequence information gleaned from an analysis of the spectrum. The class of *de-novo sequencing* algorithms attempt to reconstruct the entire peptide sequence via interpretation, without the use of a peptide database. (See for example [2, 3, 6, 9, 10, 18, 30]). To understand how this is done, consider the simple case when all observed fragments are b -type prefix ions. If all the fragment ions were present, then the sequence could be read simply by reading the ladder of increasing masses, and assigning residues to mass differences between adjacent peaks. The complexity comes from the fact that (a) there are multiple fragment ion types, including internal ions and different charge states (b) the prefix ions cannot *a priori* be separated from the suffix ions, and (c) not all positions along the peptide chain fragment.

A key algorithmic idea that deals with mixed prefix and suffix ions is the prefix residue graph, first described in Dancik et al. [9], but implicit in [3, 30]. If we knew the ion type for a spectral peak, it would uniquely define the residue mass of a prefix (PRM) of the peptide. Thus for all possible interpretations of a peak, a node

labeled with the PRM value is added to the graph and an directed edge is drawn from node u to node v if $\text{PRM}[v] - \text{PRM}[u]$ corresponds to a residue mass (adjacent fragments), or if it is close to 0 (identical fragments). Each edge is labeled with the ϵ , or the appropriate amino acid. Thus any path in this graph corresponds to a *de novo* sequence interpretation simply by concatenating the non- ϵ edge labels. A multitude of paths are possible, and scoring and ranking these paths is critical to correct interpretation. One argument against this approach is that a peak may be used multiple times in a path with different interpretations. Resolving this problem in a general scenario is equivalent to constructing paths where certain pairs of nodes are forbidden, which is known to be computationally hard [16]. However, Chen et al. [6] made the observation that the forbidden suffix-prefix pairs for tandem MS are non-intersecting. This observation allowed them to give a dynamic programming solution for finding paths with forbidden pairs. Bafna and Edwards [2] extended this approach to include multiple ion types (a, b, y , and all the neutral losses) in the interpretation. While pure *de novo* sequencing of spectra has seen many improvements, tandem mass spectra usually do not have enough information to make an unambiguous identification. Searching a database of candidate peptides (as described in the next section) constrains the possibilities enough to make such identification possible. However, *de novo* interpretation is still useful in some situations. It can be used to generate sequence tags which can then be used as additional filters to improve database search [22, 29, 30], especially for modified peptides. Also, for pure proteins whose sequence is not yet available in databases, a *de novo shotgun* sequencing of overlapping peptides obtained via multiple enzymatic digestion holds promise [21].

Scoring spectra against peptides

When the protein sequence of interest is present in a database, one can interpret an MS/MS spectrum by computing the correlation between the spectrum and a hypothetical spectrum of each peptide. The so-called *database* searching algorithms [1, 14, 15, 22, 25, 28] rely on this technique for interpretation, and have been extremely successful. Sequest [14] is the prototypic database search method. It presents a model for generation of hypothetical spectra, and a correlation function for scoring. If the spectrum is of poor quality, there is no guarantee that the top scoring peptide is the correct interpretation. Subsequent algorithms [1, 21, 24] therefore included a p -value along with the raw score to give a probability of that score arising from a randomly chosen peptide. Further improvements in scoring have come from an analysis of the physico-chemical rules of fragmentation, such as “neutral losses are more likely in the presence of acidic or basic residues”, and “proline directed fragmentation”. The algorithm in Scope [1] presents a model for quantifying these rules as probabilities, and efficient scoring with the probability functions. Recent work is directed towards data-mining and learning techniques to optimize a score function, as well as the use of intensity values in scoring [13].

Filtering

The goal of filtering is to scan a database of peptide candidates and quickly filter out the vast majority of them while retaining the true peptides for detailed scoring. While most algorithms include simple filters typified by parent mass, immonium ions, matching peaks, this topic was not actively researched until recently. With exponential growth in the number of spectra and sequence databases, this is only now beginning to see active research. Pre-indexing sequence databases is useful in removing redundant peptide information and efficient search for candidates. One approach to indexing is the use of suffix trees [12, 19]. Other approaches include the use of sequence tags as filters in a database search [22, 29, 30].

Differential analysis of protein mixtures

Differential analysis of proteome expression levels has been developing rapidly over the last years. The current standard separation technique in proteomics is undoubtedly gel electrophoresis, which is typically used together with mass spectrometry for identification of the proteins separated on the gel. While this technique is well-established and in use in most major proteomics facilities, it has disadvantages in the context of high throughput settings. In particular the difficult handling of the gels prompted several proteomics facilities (mostly in commercial settings [11], but also in academia [20, 26]) to use HPLC/MS-based techniques instead. In these approaches the liquid-chromatographic separation in an HPLC column replaces the separation on the gel. In contrast to gels, HPLC systems allow for direct coupling to a mass spectrometer, thus greatly simplifying automation of the whole analytical process. In the shotgun approach the proteins are usually fully digested in order to simplify and unify sample preparation. The combined use of the complementary expression techniques (mRNA expression, gel-based MS, and HPLC based MS) will yield further progress in diagnostics and systems biology.

The analysis of HPLC/MS data for differential analysis of protein expression poses many challenging computational problems. Note that measuring protein expression is generally difficult using MS techniques. Different

peptides have different capacity to retain charge, and so peptides with identical concentration might show up with very different intensities in the spectrum. However, the relative intensity of peaks for the *same* peptide is predictive of relative expression of that peptide in different samples. To use this fact in differential analysis, consider the output of an HPLC/MS run. As the peptides elute off the column, MS spectra are acquired in real time. The data can be represented as a 2 dimensional spectrogram, or map, with the two dimensions being LC Retention time (RT), and M/Z. As a peptide typically elutes over a fixed time span, and has a fixed M/Z, a 'spot' on this map corresponds to the elution of a peptide. The intensity of the spectra provide a third dimension. Two maps from different samples (normal and diseased) will have similar spots corresponding to the commonly expressed peptides, and spots of differential intensity corresponding to proteins whose expression levels have changed. Geometric matching algorithms are used to match the similar spots, and their intensities are used for normalization, sometimes with internal standards. Once this is done, the relative intensities of differentially expressed peptides can be computed. This basic idea can be made to be statistically robust by choosing multiple samples from both categories. This is analogous to the use of 2D gels for separation with important differences. It is the peptides, not proteins, that are being separated. As the retention time and mass measurements are typically accurate, the reproducibility of maps is much higher than 2D gels. Algorithms for creating and comparing various 2D maps, and for computing relative intensities of a few thousand differentially expressed peptides have been used to successfully identify differentially expressed peptides in tumor cells [11]. The key components of these algorithms include the de-convolution of peaks from different but overlapping peptides, and creation, normalization and comparison of multiple maps to identify differential expression.

An alternative approach uses some type of differential mass labeling of the two samples to provide uniform experimental conditions. The two samples are labeled differently, e.g. by introducing stable isotopes like ^{15}N , ^2H , ^{13}C , or ^{15}O into the proteins of one of the samples [23, 4]. Other approaches are mass-coded abundance tags (MCAT [5]) and isotope-coded affinity tags (ICAT, [17]), where the mass difference is introduced by derivatisation of specific amino acid residues. The samples are then mixed and subjected to HPLC/MS. In the resulting maps, each peptide should occur in paired 'spots' which are very similar in Retention Time, and have a mass offset that corresponds exactly to the differential mass of the labeling tags. This simplifies the computation somewhat as comparison of two LC/MS maps is not required. For statistical robustness, one might still need to perform multiple such experiments and compare maps in order to identify differentially expressed peptides unambiguously [11].

Protein structure: Cross-linking

MS technologies are clearly impacting protein identification and quantification. Can we also use them for protein structure determination? X-ray crystallography and NMR remain the dominant techniques in this field, but much work remains to be done. These techniques require large amounts of pure analyte, and even if this is available it can take many months until the structure is determined. On the other hand, it is theoretically possible to determine the tertiary structure of a protein computationally, given enough interatomic distance information [8]. Even partial information often proves to be helpful. To acquire this distance information the mass of *cross-linked* peptides can be measured [32] and using this information the sequence of the cross-linked complex can be determined. Cross-linking is a method in which certain molecules are used to specifically link two peptides. The identification of cross-linked peptides in a folded molecule then provide constraints on the spatial distance of the peptides in a folded state. One approach to identifying cross-linked peptides, is simply to find all pairs of peptides whose mass sum (plus the mass of the linker) equals the mass of the parent cross-linked molecules. Among these, a peptide pair is chosen whose theoretical spectrum best correlates with the measured spectrum([7]). These additional constraints can be used to improve tertiary structure prediction.

Conclusion

We conclude by reiterating that Mass spectrometric techniques are key to Proteomic explorations, and computational algorithms for analyzing MS data are crucial to further development of this technology. Mass spectrometry is a dynamic and evolving field and it is likely that many of the data sets and applications described here will be outdated in the coming years. Nevertheless, it is our hope that the basic understanding of MS principles and key algorithmic components will continue to be useful for future proteomic applications of mass spectrometry.

References

- [1] V. Bafna and N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17 Suppl 1:S13–21, June 2001. Appeared in Intl. Conference on Intelligent Systems for Molecular Biology.

- [2] V. Bafna and N. Edwards. On *de novo* interpretation of peptide spectra. In *International Conference on Computational Molecular Biology (RECOMB)*, pages 9–18, 2003.
- [3] C. Bartels. Fast algorithm for peptide sequencing by mass spectrometry. *Biomedical and Environmental Mass Spectrometry*, 19:363–368, 1990.
- [4] S.J. Berger, S.-W. Lee, G.A. Anderson, L. Pasa-Tolic, N. Tolic, Y. Shen, R. Zhao, and R.D. Smith. High-throughput global peptide proteomic analysis by combining stable isotope amino acid labeling and data-dependent multiplexed-MS/MS. *Anal. Chem.*, 74:4994–5000, 2002.
- [5] G. Cagney and A. Emili. De novo peptide sequencing and quantitative profiling of complex protein mixtures mass-coded abundance tagging. *Nature Biotechnol.*, 20:163–170, 2002.
- [6] 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. *Journal of Computational Biology*, 8(6):571–83, 2001.
- [7] Ting Chen, Jacob D. Jaffe, and George M. Church. Algorithms for identifying protein cross-links via tandem mass spectrometry. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB01)*, pages 95–102, 2001.
- [8] F.E. Cohen and M.J. Sternberg. The use of chemically derived distance constraints in the prediction of protein structure with myoglobin as an example. *Journal of Molecular Biology*, 137:9–22, 1980.
- [9] V. Dancik, T. Addona, K. Clauser, J. Vath, and P.A. Pevzner. De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6:327–342, 1999.
- [10] J. Fernández de Cossio, J. Gonzales, and V. Besada. Protein identification using mass spectrometric information. *Comput. Appl. Biosci.*, 11:427–434, 1995.
- [11] Bruno Domon, Kim Alving, Tao He, Terence E. Ryan, and Scott D. Patterson. Enabling parallel protein analysis through mass spectrometry. *Curr. Opin. Mol. Therapeutics*, 4:577–586, 2002.
- [12] Nathan Edwards and Ross Lippert. Generating Peptide Candidates from Amino-Acid Sequence Databases for Protein Identification via Mass Spectrometry. In *Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, pages 68–81. Springer-Verlag, 2002.
- [13] J.E. Elias, F.D. Gibbons, O.D. King, F.P. Roth, and S.P. Gygi. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nature Biotechnology*, 22:214–219, 2004.
- [14] J. Eng, A. McCormack, and J. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of American Society of Mass Spectrometry*, 5:976–989, 1994.
- [15] D. Fenyo, J. Qin, and B.T. Chait. Protein identification using mass spectrometric information. *Electrophoresis*, 19(6):998–1005, 1998.
- [16] M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H. Freeman and Company, 1979.
- [17] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.*, 17:994–999, 1999.
- [18] R.J. Johnson and K. Biemann. Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomedical and Environmental Mass Spectrometry*, 18:945–957, 1989.
- [19] B. Lu and T. Chen. A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion and post-translational modification. *Bioinformatics*, 2003.
- [20] Michael J. MacCoss, Christine C. Wu, Hongbin Liu, Rovshan Sadygov, and John R. Yates III. A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. *Anal. Chem.*, 2003.
- [21] M.J. MacCoss, W.H. McDonals, A. Saraf, R. Sadygov, J.M. Clark, J.J. Tasto, K.L. Gould, D. Wolters, M. Washburn, A. Weiss, J.I. Clark, and J.R. Yates. Shotgun identification of protein modifications from protein complexes and lens tissues. *Proceedings of the National Academy of Sciences*, 99(12):7900–7905, 2002.

- [22] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66:4390–4399, 1994.
- [23] Y. Oda, K. Huang, F. R. Cross, D. Cowburn, and B. T. Chait. Accurate quantitation of protein expression and site-specific phosphorylation. *Proc. Natl. Acad. Sci. USA*, 96:6591–6596, 1999.
- [24] D.N. Perkins, D.J. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551–3567, 1999.
- [25] P.A. Pevzner, V. Dancik, and C.L. Tang. Mutation-tolerant protein identification by mass-spectrometry. In R. Shamir, S. Miyano, S. Istrail, P.A. Pevzner, and M.S. Waterman, editors, *International Conference on Computational Molecular Biology (RECOMB)*, pages 231–236. ACM Press, 2000.
- [26] Albert Sickmann, Jörg Reinders, Y. Wagner, C. Joppich, et al. The proteome of *saccharomyces cerevisiae* mitochondria. *Proc. Natl. Acad. Sci. USA*, 100:13207–12, 2003.
- [27] G. Siuzdak. *Mass Spectrometry for Biotechnology*. Academic Press, 1996.
- [28] D. L. Tabb, J. K. Eng, and J. R. 3rd Yates. *Protein Identification by SEQUEST*, volume 1, pages 125–142. Springer, 2001.
- [29] D.L. Tabb, A. Saraf, and J.R. Yates. GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Analytical Chemistry*, 75:6415–6421, 2003.
- [30] J.A. Taylor and R.S. Johnson. Sequence database searches via *de novo* peptide sequencing by mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11:1067–1075, 1997.
- [31] M.R. Wilkins, K.L. Williams, R.D. Appel, and D.F. Hochstrasser. *Proteome Research: New Frontiers in Functional Genomics*. Springer Verlag, 1997.
- [32] Malin Young, Ning Tang, Judith C. Hempel, Connie M. Oshiro, Eric W. Taylor, and Irwin D. Kuntz. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proceedings of the National Academy of Sciences*, 97:5802–5806, 1997.