De Novo Peptide Sequencing via Tandem Mass Spectrometry and Propositional Satisfiability

Renato Bruni
bruni@diei.unipg.it or bruni@dis.uniroma1.it
University of Perugia

I FIMA International Conference "Models and Methods for Human Genomics"
January 23-27, 2006, Champoluc, AO, Italy

Outline

- Peptide Analysis via Tandem Mass Spectrometry
- From the MS/MS Spectrum to the Sequence
- The Peak Interpretation Problem as Satisfiability
- The Actual Sequencing Phase
- Computational Complexity
- Results

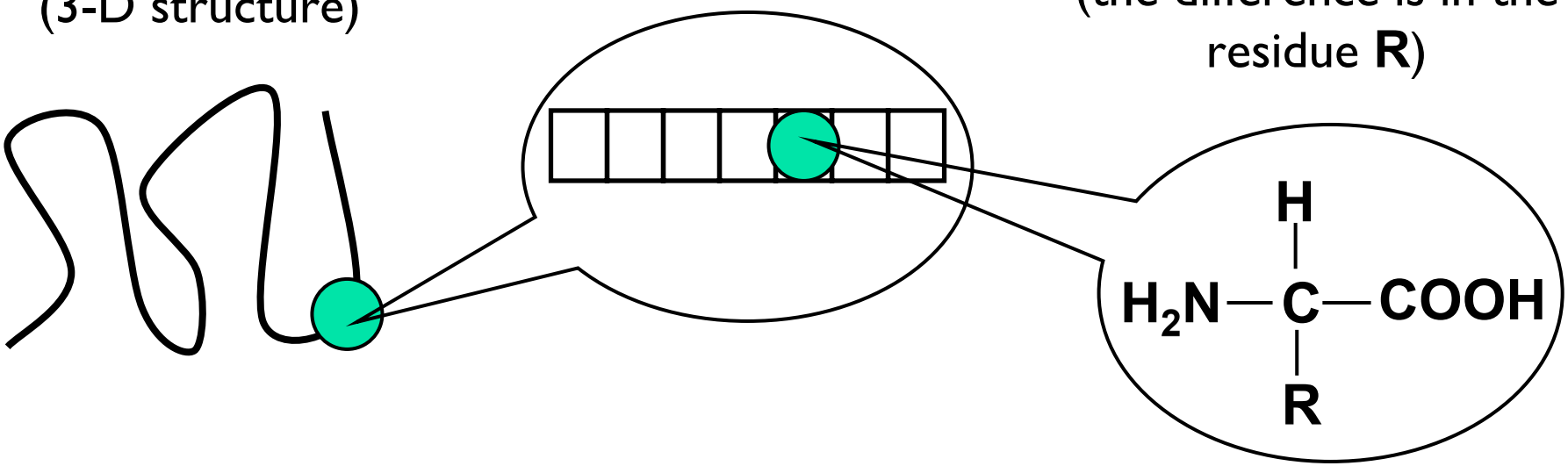
Proteins, Peptides & Aminoacids

- **Peptide analysis** is one of the most important issues in biological and medical research: **proteins** are made of **peptides**
- After the **Genome Project**, the **Proteome Project** started

Protein
(3-D structure)

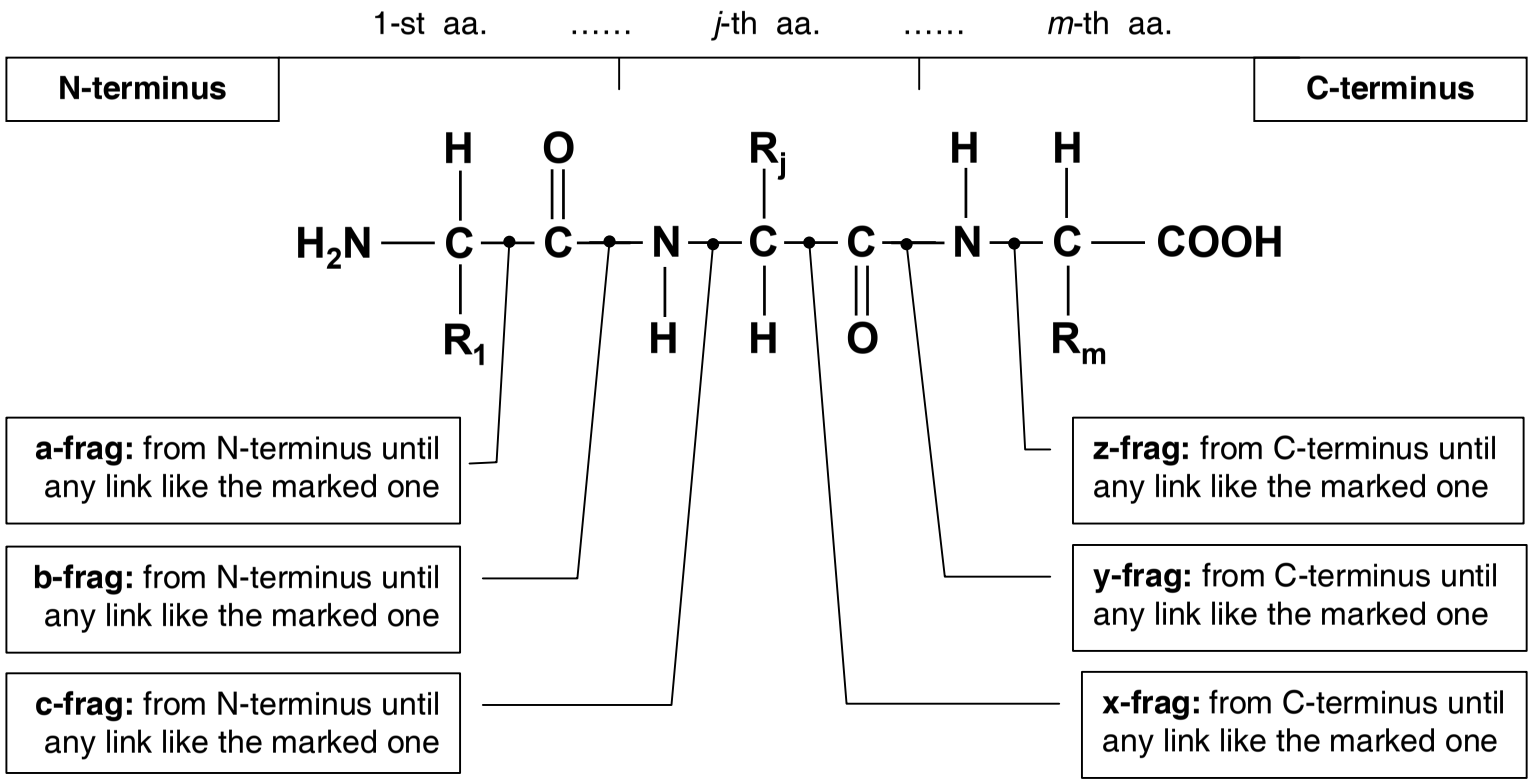
Peptide
(sequence of aminoacids)

Aminoacid
(the difference is in the residue **R**)



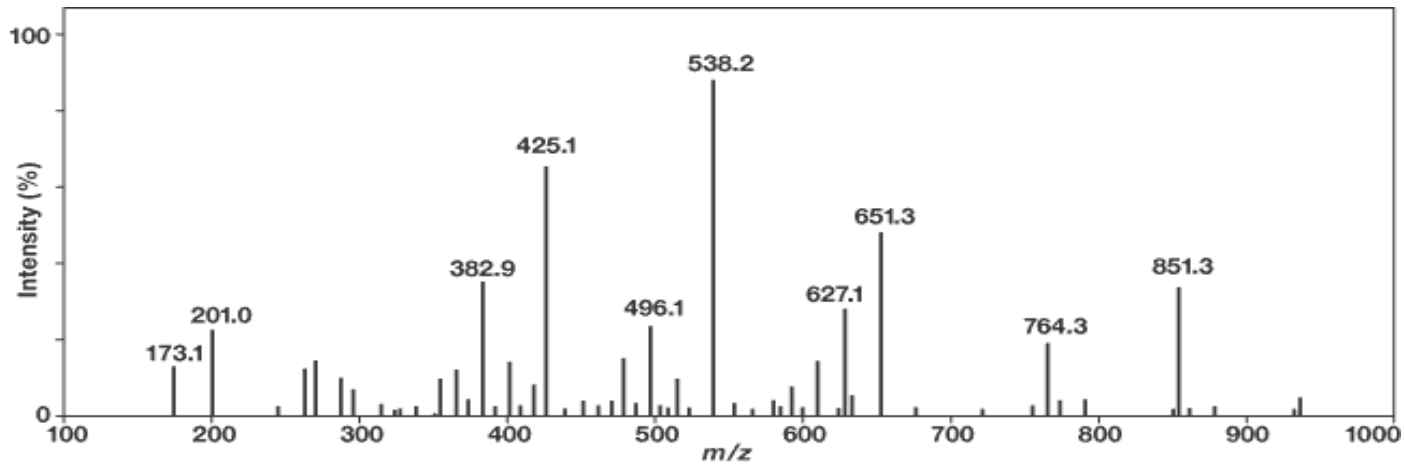
De Novo Peptide Analysis via MS/MS

- A small number of molecules of a peptide (the **compound**) undergo **tandem mass spectrometry** (MS/MS)
- Most of such molecules **break**. Fragmentation is a **stochastic** process, but most fragments are of **standard** known types



The MS/MS Spectrum

- We obtain an MS/MS **spectrum**: for each molecular weight *w*, the amount of material having weight *w* (**peak**)



- We know molecular **weights** of aminoacids (the **components**) and rules for computing the weights of standard **fragments** (unfortunately, **not simply the sum** of the components)
- Thanks to such fragmentation, we can in principle find which **sequence** (or sequences) of components could produce a **spectrum** like **ours**

Prior Work I

- This problem is hard. Showed **NP-complete** [e.g. Bafna-Edwards, 2003] in the general case. Has a strong combinatorial nature
- Until few years ago biochemists used a **trial-by-hand** approach and/or a **database** approach [e.g. Johnson-Taylor, 2000]. But the number of possible peptides is **huge**
- Recently, (i) looking for **continuous series** of fragments differing by just one aminoacid, or (ii) generating **random sequences** and check how they fit the spectrum [software developed by mass spectrometry producers: DeNovoX, MassSeq, Peaks]. But fragmentation is **never complete** and **completely standard**, and number of possible sequences is **huge**

Prior Work II

- **Graph theoretical** construction [Dancik-Addona-Clauser-Vath-Pevzner, 1999] and **dynamic programming** approach [Chen-Kao-Tepel-Rush-Church, 2001; Bafna-Edwards, 2003]. Limitations in the types and charges of fragments
- Mathematical **feasibility model** with binary variables ($x_{ij} = 1$ if aminoacid i is in position j of peptide) and **branching** algorithm [Bruni-Gianfranceschi-Koch, 2004]. Still computationally expensive

Rules for Weights

- Weights of the possible **components** $A = \{a_1, \ldots, a_n\} \quad a_i \in \mathrm{R}_+$
- **Peaks** extracted from spectrum $W = \{w_0, w_1, \ldots, w_f\} \quad w_j \in \mathrm{R}_+$
- **Number of molecules** of each component in **fragment** j (the one producing peak j) $Y^j = \{y_1^j, \ldots, y_n^j\} \quad y_i^j \in \mathrm{Z}_+$
- **Observed weight** of fragment j

$$w_j = \frac{\sum_{i \in N} [\, y_i^j (a_i - 18.015)\,] + c}{e} \pm \delta$$

where e is the number of **charges** retained by the fragment,

δ the maximum numerical **error** of spectrometer,

and c is -26.994 for **a-frag**, 1.008 for **b-frag**, 18.039 for **c-frag**, 45.017 for **x-frag**, 19.023 for **y-frag**, 3.000 for **z-frag**

The Peak Interpretation Problem

- So, the **interpretation** of each peak is **determinant** !
- If a peak of weight w is considered for instance a-frag, it may have a **certain sequence**, if is b-frag it **cannot** have that sequence
- We **do not know** from the spectrum the **type** of fragment which originated each peak
- We isolate **peak interpretation** as a problem in itself
- In practice, many other minor **complications** (non standard fragments, isotopes, numerical precision, experimental errors, impurities, noise, etc.)

A Logic-Based Approach

- We define **peak interpretation**: assigning to each peak j in the spectrum one hypothesis about **type** $t \in \{a, b, c, x, y, z\}$ and **charge** $e \in \{1, \ldots, e_{\max}\}$ of its originating fragment
- All interpretations given to all peaks must be **coherent**, i.e. respect some rules given by incompatibilities and multicharges (formalized later)
- We express all this by using **propositional logic**:

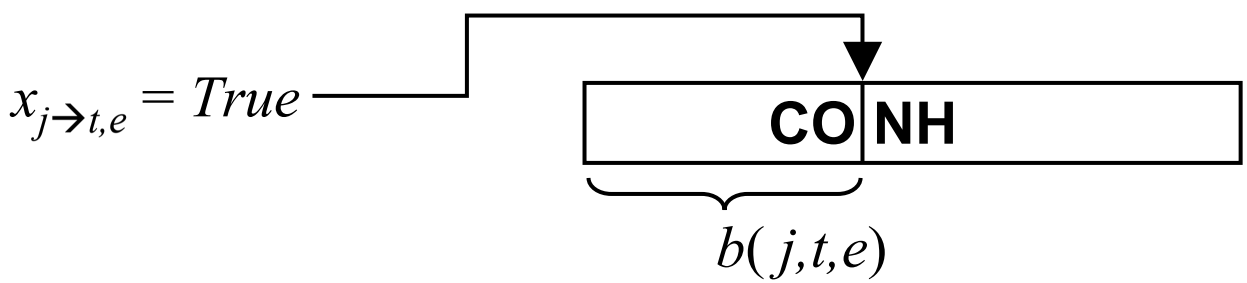
$$x_{j \to t,e} \in \{True, False\}$$

Example: "peak j must be one of the 6 standard types and have 1 or 2 charges" corresponds to the clause

$$(x_{j \to a,1} \vee x_{j \to b,1} \vee x_{j \to c,1} \vee x_{j \to x,1} \vee x_{j \to y,1} \vee x_{j \to z,1} \vee x_{j \to a,2} \vee \ldots \vee x_{j \to z,2})$$

The Normalized Peptide

- In order to work better, we define a theoretical model of peptide, the **normalized peptide**. Its weight, and the weights of its fragments, are simply the sum of those of its components
- Each **real** peptide has its corresponding **normalized** peptide. Weight transformation rules may be found
- **N-terminal portion** of a normalized peptide: sequence of components going from **N-terminus** until a **bond** between two components
- Now, **variable** $x_{j \to t,e} = True$ **implies** the existence of an N-terminal (=left) portion of normalized peptide of **weight** $b(j,t,e)$



Writing Our Knowledge into Clauses

- For each peak, we need a clause for its **interpretation**
- If two variables produce N-terminal portions having a difference b'- b'' which cannot be any sequence of normalized components (e.g. they differ of 1), they are **incompatible**: they cannot be both true. This gives the incompatibility clause

$$(\neg\, x_{b'} \vee \neg\, x_{b''})$$

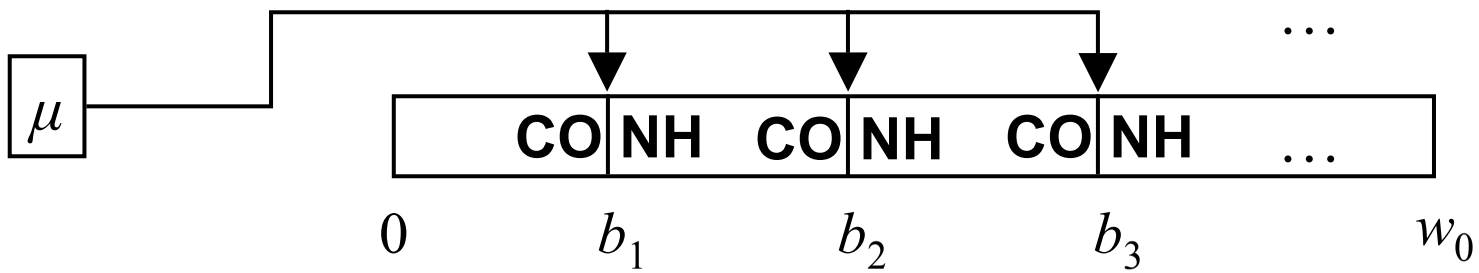
- The presence of **multicharged** fragments usually implies that of the corresponding **monocharged** ones. This gives the clause

$$(\, x_{j' \to t,e} \rightarrow x_{j'' \to t,1}\,) = (\, \neg\, x_{j' \to t,e} \vee x_{j'' \to t,1}\,)$$

- Other problem specific clauses may be written

Models of the Formula

- We have a **CNF logic formula** F encoding our knowledge. Each **model** μ (i.e. variable assignment making $F = True$) is a coherent **interpretation** of the set W of peaks extracted from spectrum
- Each model determines the **location of bonds** in the (unknown) peptide under analysis



- Finding a model of a CNF formula is the **Satisfiability** problem (SAT). **NP-complete** but **very well studied** !
- We want **all** possible solutions of the peptide analysis, hence **all models** $\{\mu_1, \ldots, \mu_r\}$ of F

The SAT Solver

- The SAT solver **BrChaff** is a **complete deterministic** satisfiability solver [Bruni, 2004]. Belongs to the family of **DPLL** solvers
- Like most advanced SAT solvers, uses **preprocessing** and **non-chronological backtracking** and **conflict based learning**
- Has an original branching rule called **Reverse Assignment Sequence** (RAS), based on the **history of the conflicts** obtained during the search. Tries to avoid unpleasant situations in the exploration of the branching tree
- Obtained **good** results in the SAT Competition 2004 (Vancouver, Canada)
- With a slight modification for finding **all models**

Generation of the Subsequences

- After locating the bonds, we need to **sequence** each portion between two adjacent bonds (b_k , b_{k+1}), i.e. finding **all** possible sequences of components having **weight** $b_{k+1} - b_k$
- We compute all **non-negative integer vectors** (y_1 ,…, y_n)' verifying

$$b_{k+1} - b_k = y_1 (a_1 - 18.015) + \ldots + y_n (a_n - 18.015) \pm 2\delta$$

- We use a specialized **branching** algorithm, but dynamic programming could as well be adopted
- Sequencing each portion is **much easier** than sequencing the whole
- This is possible thanks to our peak **interpretation**, hence to our hypothesis on the **location** of the bonds

Concatenation of the Subsequences

- After we have a set of subsequences $S(b_{k+1} - b_k)$ for each portion of the peptide, we simply make their **concatenation** in all possible ways
- All the **sequences** S_μ corresponding to **model** μ are

$$S_\mu = S(b_1 - 0) \oplus S(b_2 - b_1) \oplus \ldots \oplus S(w_0 - b_p)$$

- Finally, **all the possible sequences** compatible with the given spectrum are obtained by taking all those corresponding to each model of the CNF formula

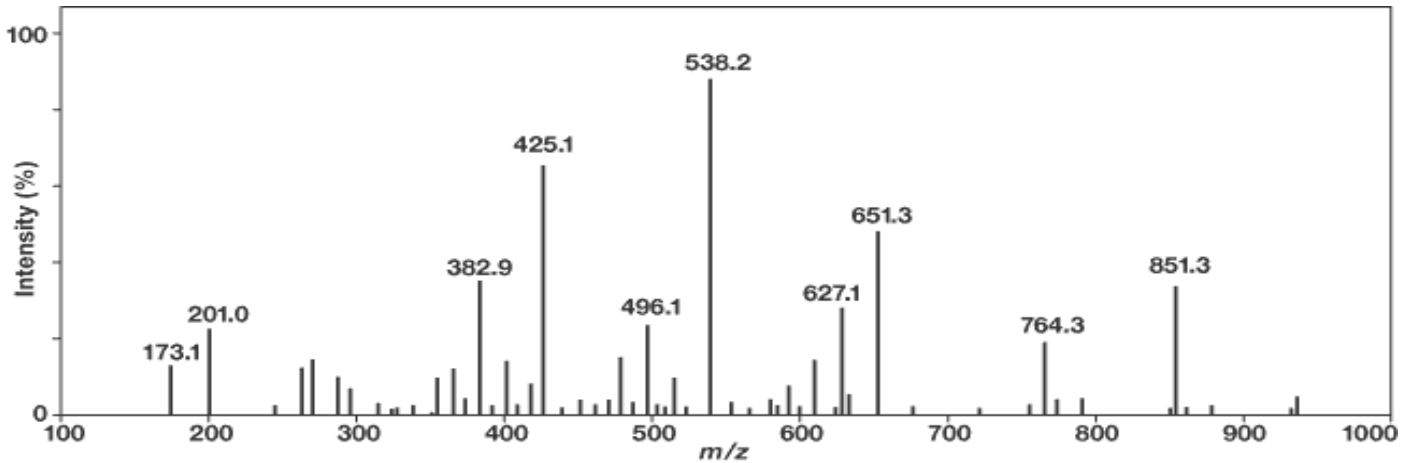
$$S = S_{\mu_1} \cup S_{\mu_2} \cup \ldots \cup S_{\mu_r}$$

- We can prove that this are **all and only** the solutions compatible with our selected spectrum

Why so many Solutions?

- Since we analyze **one** peptide, we would like to obtain **one** solution
- The problem is the quantity on **information** contained in the spectrum
- If we have enough information in the spectrum, we have a **unique** solution
- If there is **not enough** information, we have **more than one** possible solution
- If there are **problems** in the spectrum, we may have **no** solutions (or, obviously, wrong solutions)
- Since there are **often** problems in the spectrum (experimental errors, non standard fragmentation, noise, etc.), when interpreting all peaks is impossible, we use the models $\{\mu_1, \ldots, \mu_r\}$ interpreting the **most** of them

Example: Spectrum → Interpretations



- We allow 20 possible **aminoacids**
- We allow **fragments** a, b, c, x, y, z and 1 or 2 **charges**
- We extract w_0=851.3, w_1=764.3, w_2=651.3, w_3=627.1, w_4=538.2, w_5=496.1, w_5=425.1, w_6=382.9, w_7=201.0, w_8=173.1
- We obtain a **formula** with 108 var and 4909 clauses having **3 models**: 3 possible coherent interpretations for these peaks

Example: Interpretations → Sequences

1. Bonds in { 87.0, 224.2, 339.2, 452.2, 565.2, 662.2} → two **sequences**:
   Ser-His-Asp-Leu-Leu-Pro-Gly-Leu
   Ser-His-Asp-Leu-Leu-Pro-Leu-Gly
2. Bonds in { 87.0, 224.2, 339.2, 452.2, 565.2, 678.3} → two **sequences**:
   Ser-His-Asp-Leu-Leu-Leu-Gly-Pro
   Ser-His-Asp-Leu-Leu-Leu-Pro-Gly
3. Bonds in { 87.0, 184.0, 355.2, 452.2, 565.2, 662.2} → four **sequences**:
   Ser-Pro-Gly-Asn-Pro-Leu-Pro-Gly-Leu
   Ser-Pro-Gly-Asn-Pro-Leu-Pro-Leu-Gly
   Ser-Pro-Asn-Gly-Pro-Leu-Pro-Gly-Leu
   Ser-Pro-Asn-Gly-Pro-Leu-Pro-Leu-Gly

- Even if no unique solution, they have **similar** structure
- If **presence** or **absence** of some component is known, **less** possibilities

Computational Complexity

- A peptide with n aminoacids may have (with a reasonable collision energy) **~30%** of **unbroken** bonds. The number of **solutions** is **~2^(0.3n)**. Clearly, worst case is no broken bonds (n! solutions, all the permutations) but that would be nonsense. The number of **broken** bonds may be **~70%**. Each bound may break in three points, each fragment may appear (or not) in the spectrum. We have **~6/2(0.7n) peaks**.
- Therefore, Chen *et al*., 2001 find a solution in O(n + n^2) time, and all solutions in O(n * 2^n + n^2), and Bafna-Edwards, 2003 find the most probable interpretation in O(n^3 * logn). They use some restrictions and find **one solution** in **polynomial** time, but **all solutions** in **exponential** time.
- Our approach solves SAT, which is polynomially solvable under special conditions (e.g. only b and y fragment, or other restrictions), although exponential if the problem without restriction is considered. So, finding one solution for the **restricted** problem is **polynomial**, for the **full** problem is **exponential**, finding **all solutions** is **exponential**.

Implementation

All the procedure have been implemented in C++. It takes in input:

- List of **components** (any)
- **Spectrum** of compound (either a full spectrum, selecting higher peaks according to some parameters, or a list of manually selected peaks)
- Types and rules for **fragments** and **charges**
- Generates and solve the **SAT instance** encoding the peaks interpretation problem
- For each **model**, finds and outputs the set of all **sequences**

A demo version of the software Peptide Analyzer 2005 is downloadable from www.polydart.com

Real World Peptide Sequencing

On a **normal Pentium** IV 2.4GHz with 1Gb RAM

Weight (D.)	# Var	# Cla	# Mod	# Seq	Time (sec.)
572.2	84	3571	2	2	1.7
851.3	144	6780	4	7	2.0
859.1	240	8156	5	29	4.1
913.3	288	10741	8	32	6.8
968.5	228	7021	10	38	4.1
1023.6	262	5564	15	40	6.1
1108.6	168	7456	16	64	12.2
1479.8	40	690	7	22	14.3
1570.6	264	9657	14	98	56.8

Thanks to people who provided the mass spectra

Conclusions

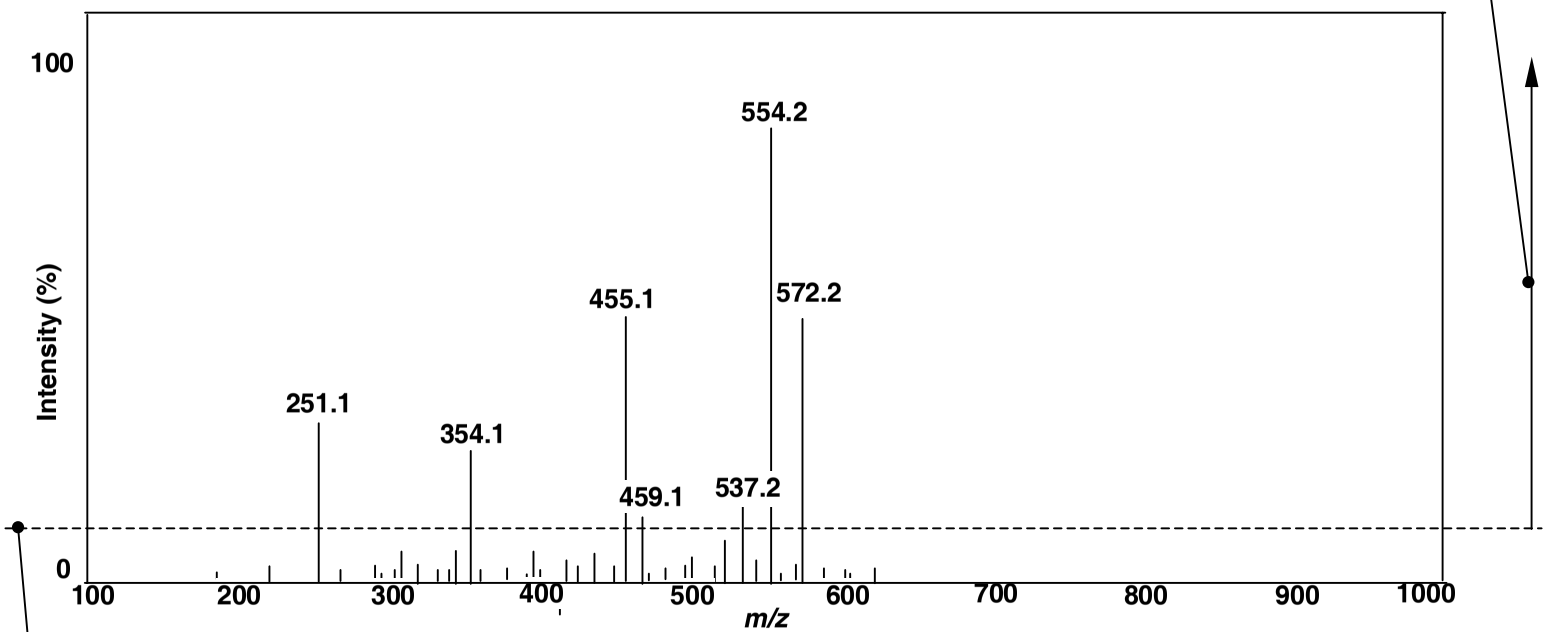
- De Novo Peptide Sequencing is a very **relevant** and **difficult** problem
- The proposed approach **guarantees** to find **all solutions** compatible with a given selected spectrum in very reduced computational times
- Since **spectrometry** itself may suffer from several problems, a unique solution is **seldom** obtainable. E.g., if a part of the peptide did not break, it is **impossible** to exactly know that part of the sequence (we may know that contains A,B,C but in which order?)
- However, the obtained solutions are generally very related
- For all **synthetic** peptides (for which we know the sequence) analyzed up to now, the set of obtained sequences **included** the **real one** !

Acknowledgements

- Thanks to Prof. G. Gianfranceschi and Prof. G. Koch for traveling together the long research path necessary for clarifying various aspect of the fragmentation process
- Thanks to Prof. L. De Angelis and Mr. F. Giavarini for the experimental mass spectrometry analyses.
- Thanks to Ing. A. Moscatelli and Ing. A. Santori for helping the implementation work

EXAMPLE

Considered values: all values above limit



Limit: line delimiting 10% of the maximum intensity

Solution obtained: H—**L**—**H**—**C**—**T**—**V**—OH