Development of Complex Curricula for Molecular Bionics and Infobionics Programs within a consortial* framework**

Consortium leader

PETER PAZMANY CATHOLIC UNIVERSITY

Consortium members

SEMMELWEIS UNIVERSITY, DIALOG CAMPUS PUBLISHER

The Project has been realised with the support of the European Union and has been co-financed by the European Social Fund ***

**Molekuláris bionika és Infobionika Szakok tananyagának komplex fejlesztése konzorciumi keretben

***A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

INTRODUCTION TO BIOINFORMATICS

(BEVEZETÉS A BIOINFORMATIKÁBA)

CHAPTER 7

DNA/Protein Sequencing Algorithms

(DNS és fehérje szekvenálási algoritmusok)

András Budinszky

Definitions

DNA sequencing refers to methods for determining the order of the nucleotide (adenine, guanine, cytosine, and thymine) in a DNA molecule.

Protein sequencing refers to methods for identifying the amino acid sequence of a protein.

These two processes are playing a central role in basic biological research and in numerous applied fields such as diagnostic, biotechnology, system biology, drug development, forensic biology, etc.

History of DNA Sequencing

1953 Discovery of the structure of the DNA double helix (Watson & Crick)

1972 Development of recombinant DNA technology (permits isolation of defined fragments of DNA

1972-6 Sequence of the first complete gene and the complete genome of bacteriophage MS2 (Friers)

1977 Sequencing by chemical degradation (Gilbert)

Sequencing with chain-terminating inhibitors (Sanger)

1984 Decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.

History of DNA Sequencing (cont)

1987 Marketing the first automated sequencing machine (Applied Biosystems)

1988 Sequencing by hybridization suggested as an alternative sequencing method

Sequencing with chain-terminating inhibitors (Sanger)

1991 Sequencing of human expressed sequence tags (ESTs) begins (Craig Verter's lab).

Light directed polymer synthesis developed (Steve Fodor)

1994 Affymetrix develops first 64-kb DNA microarray

History of DNA Sequencing (cont)

1995 Publish the first complete genome of a free-living organism (bacterium Haemophilus influenzae, 1,830,137 bases, Craig Venter)

1996 Introducing pyrosequencing (sequencing by synthesis, Nyren)

2001 A draft sequence of the human genome is published (Nature, Science)

2004 Markets a parallelized version of pyrosequencing machine (454 Life Sciences, first version reduced costs 6-fold compared to automated Sanger sequencing)

Shotgun Sequencing

A method used for sequencing long DNA strands.

Sequences are randomly subdivided into millions of smaller fragments by cutting with restriction enzymes or by shearing with mechanical forces.

About the first 500 – 700 nucleotides from each small fragments are sequenced ("read") by the Sanger method.

Multiple overlapping reads are obtained by performing several rounds of this fragmentation and sequencing.

Finally, the overlapping ends of different reads are assembled by computer program into a continuous sequence (see next slides)

Assembling the Fragments

In computational problem sense, the assembly task can be defined as the Shortest Superstring Problem (SSP).

SSP is looking for the shortest string which contains each member of a given set of strings.

SSP has relevance in other areas such as data compression, sparse matrix compression.

A greedy algorithm (finds only an approximation)

- picks those two strings that overlap in the most characters
- merges them
- repeats this until only one string left (a superstring).

Shortest Superstring Problem (SSP)

Unfortunately, finding the optimal solution for this problem is $\mathcal{NP}$-hard (that is, the problem cannot be solved in polynomial time).

Proof:

We can show that SSP corresponds Traveling Salesman Problem (TSP), which is known to be $\mathcal{NP}$-complete.

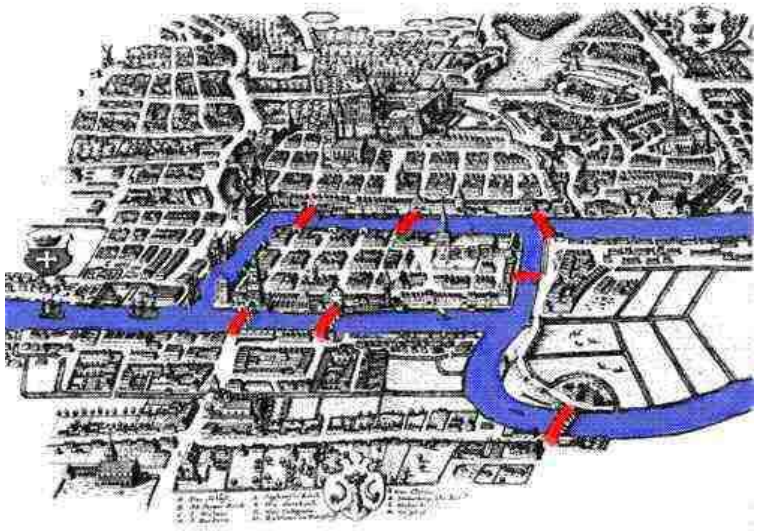
In order to facilitate the proof, we will introduce Hamiltonian cycle/path on graphs (and meanwhile we cover Eulerian cycle/path needed for another topic).

Königsberg Bridge Problem

The city of Königsberg had two islands which were connected to each other and the mainland by seven bridges.

People tried to find a way to walk all seven bridges without crossing a bridge twice.

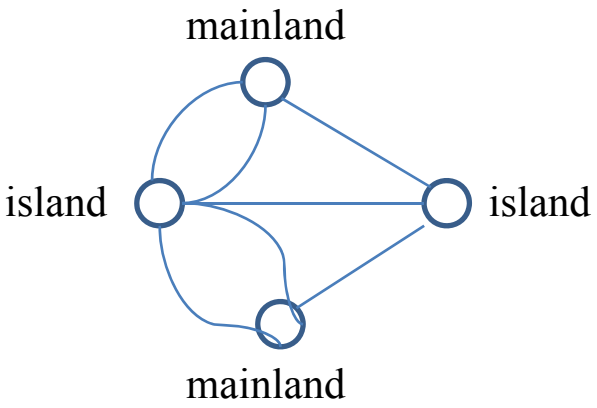
Finally, in 1735 – using a graph – Euler proved that the problem has no solution. This was actually the foundation of graph theory.



Abstract Definition of Königsberg Bridge Problem

Reformulation of the problem in abstract terms: in a graph the vertices represent the islands and the mainland and the edges stand for the bridges.

Then a path needs to be found which crosses every bridge exactly once.



Eulerian Cycle/Path

Eulerian cycle in a graph is one that visits every edge exactly once.

- Such cycle exists if and only if the degree of each vertex is even.

  Note: The degree of a vertex is the number of edges touching it.

Eulerian path in a graph is one that visits every edge but the start and end vertices do not have to be the same.

- Such cycle exists if and only if the graph contains zero or two (start and end) vertices of odd degree.

Algorithm for Finding an Eulerian Cycle

A. Starting from an arbitrary vertex "walk" along unused edges until the start vertex is reached.

B. If the Eulerian cycle has not been constructed yet, then there must be a vertex (v) along this route which has an untraversed edge. Execute step A again starting from vertex v.

C. Combine this new route with the previous one into a single cycle through vertex v.

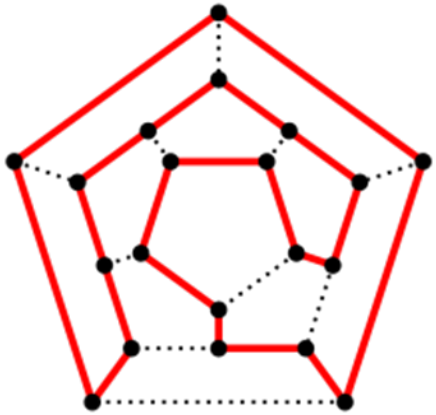
This algorithm is linear in time.

Hamiltonian Cycle/Path

Hamiltonian cycle in a graph is one that visits every vertex exactly once.

Hamiltonian path in a graph is one that visits every vertex but the start and end vertices do not have to be the same.

Unfortunately, the problem of constructing such cycle or path is $\mathcal{NP}$-complete.



Originally it was defined on a game (Icosian, in a dodecahedron) invented by Hamilton in 1857

Correspondence between SSP and TSP

Constructing a graph for SSP:

- Vertices represent the n strings $s_1, s_2, \ldots, s_n$
- Edges are drawn between such vertex pair s_i and s_j for which prefix of s_j matches suffix of s_i; the length of the edge should be equal with the number of overlapping characters.

Now SSP is to find the longest path which visits every vertex exactly once.

This is exactly the same as the Traveling Salesman Problem (shortest and longest reversed) which is also $\mathcal{NP}$-hard.

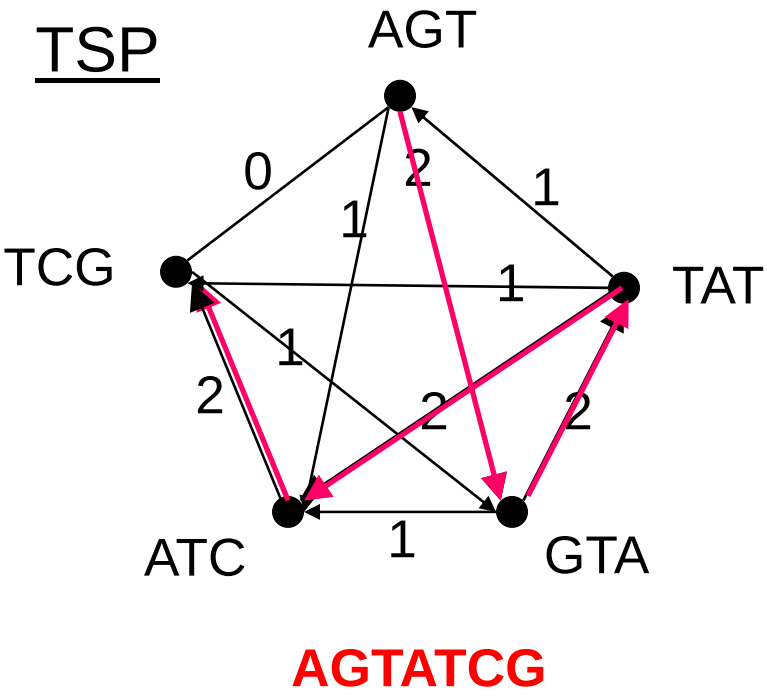
An Example for SSP to TSP

Given s = AGTATCG segment.

SSP

TCG
TAT
AGT
AGTATCG
GTA
ATC

TSP



AGTATCG

Summary of the Shotgun Sequencing

Systems using this method work in three phases:

- Overlap – Generate potentially overlapping reads. Find the best match between the suffix of one read and the prefix of another. Correct errors using multiple local alignment.
- Layout – Merge reads into contigs and those into supercontigs. Repeats are major problems.
- Consensus – Derive the DNA sequence and correct read errors.

Sequencing by Hybridization (SBH)

A non-enzymatic method that uses a DNA microarray.

The microarray is prepared by attaching all possible DNA probes of length ℓ in a systematic order.

Copies of a DNA fragment to be sequenced is fluorescently labeled.

The dyed DNA fragments are hybridized to the array.

DNA fragments hybridize with those probes that are complementary to substrings of length ℓ of the fragments.

Sequencing by Hybridization (SBH)

Using a spectroscopic detector, it is determined which probes hybridize to the DNA fragment.

In this way we get the ● -mer composition of the target DNA fragment.

Finally, apply a combinatorial algorithm to reconstruct the sequence of the target DNA fragment from the ● -mer composition (*spectrum*(*s*, *l*) – the *unordered* set of all possible *(n* – ● + *1)* ● -mers of a string *s* of length *n*).

Commercial system Affimetrix and Complete Genomics Inc. use this technology.

An Example for *spectrum*(*s, l*)

Given *s* = AGTATCG segment.

Since elements of *spectrum*(*s, l*) are unordered by definition, all of the following are equivalent representations of *spectrum (s, 3):*

{AGT, GTA, TAT, ATC, TCG}

{ATC, AGT, TAT, GTA, TCG}

{AGT, ATC, GTA, TAT, TCG}

It is customary to use the lexicographically maximal representation as the canonical one (here the 3rd one).

Note: Different sequences may have the same spectrum.

Solving SHB with a Hamiltonian Path

Constructing a directed graph (DAG) of SBH for a given *spektrum(s, ●)*:

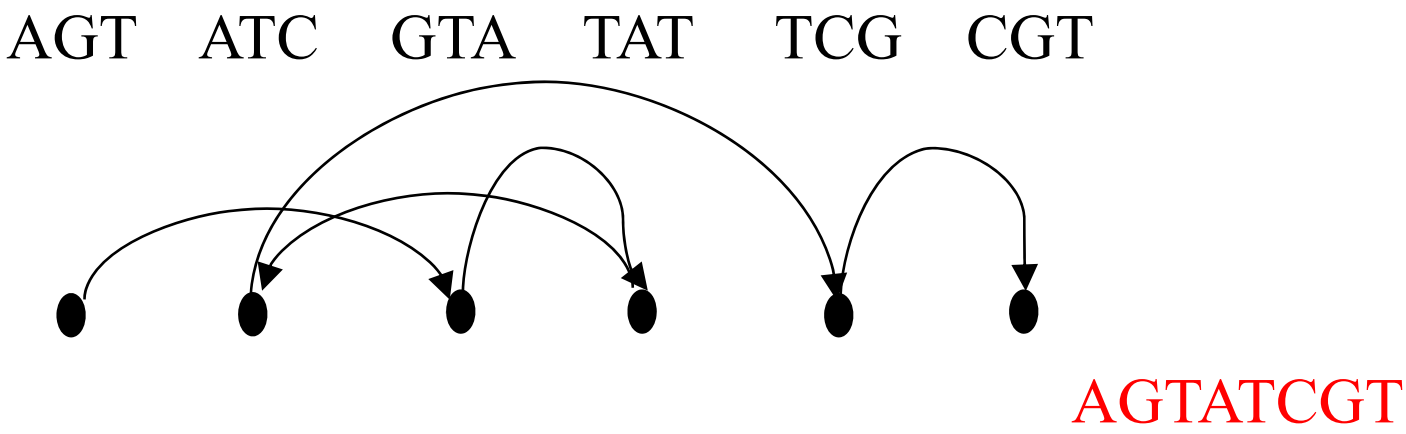
- Vertices represent the ●-mers of the *spektrum*.
- Edges are drawn between each vertex pair s_i and s_j for which prefix of s_j overlaps suffix of s_i in the length of ● - 1. This edge will be $s_i \rightarrow s_j$.

The Hamiltonian path in this graph provides the solution of the SHB problem therefore it is an $\mathcal{NP}$-hard solution.

Note: Multiple path (that is, multiple potential solution) might exist.

An Example for Hamiltonian Path Approach

Given *spectrum* = {AGT, ATC, CGT, GTA, TAT, TCG}.



Path visits every vertex once.

Note: Multiple paths (and thus solution) might exist.

Solving SHB with an Eulerian Path

Constructing a directed graph (DAG) of SBH for a given *spektrum(s, ●)*:

- Vertices represent the (*●-1*)-mers of the *spektrum*.
- Edges are drawn between each vertex pair s_i and s_j for which there exists an *●*-mer *x* such that the first *●* -1 nucleotides of *x* matches *q* and the last *●* -1 nucleotides of *x* matches *p*. This edge will be $s_i \rightarrow s_j$.

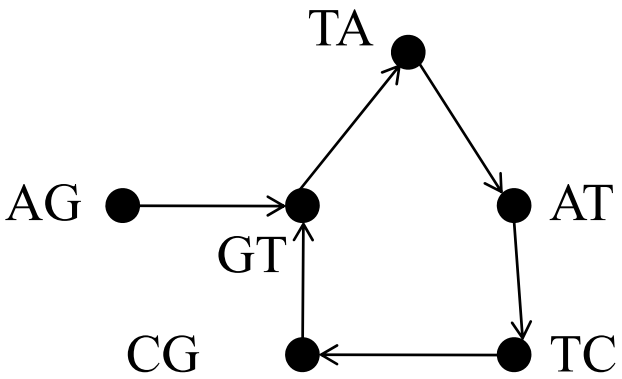
The Eulerian path in this graph provides the solution of the SHB problem therefore it is a linear solution.

An Example for Eulerian Path Approach

Given *spectrum* = {AGT, ATC, CGT, GTA, TAT, TCG}.

Vertices ((ℓ-*1*)-mers): {AG, AT, CG, GT, TA, TC}.

Edges (correspond to ℓ-mers): e.g. AG→GT belongs to AGT.



AGTATCGT

Path visits every edge once.

Difficulties with SBH

It is difficult to differentiate between probes hybridized with perfect matches and the ones with 1 or 2 mismatches.

This problem can be decreased with longer *l*-mers, but array size increases exponentially in *l*; however, array size is limited with current technology.

The bottom line is that SBH is still impractical. As DNA microarray technology improves, SBH may become practical in the future.

This technology has largely been displaced by Sequencing by Synthesis based methods.

High-throughput Sequencing Technologies

They are the so-called "next-generation" sequencing technologies.

They parallelize the sequencing and produce vast amount of sequences at once.

- Massively Parallel Signature Sequencing (MPSS), developed by Lynx Therapeutics. Later merged with Solexa and led to the development of sequencing by synthesis (see slides 28-29)
- DNA Nanoball Sequencing, developed by Complete Genomics. Short sequences of DNA are determined from each DNA nanoball and its major difficulty is the mapping of short reads to a reference genome.

More Next-Generation Technologies

- Polony Sequencing, developed in the laboratory of George Church at Harvard. It combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry.
- Illumina (Solexa), developed a technology based on reversible dye-terminators.
- SOLiD Sequencing, developed by Applied Biosystems, uses sequencing by ligation. It has incorporated Polony sequencing.

Sequencing by Synthesis

The method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step.

The template DNA is immobile and solutions of A, C, G, and T nucleotides are sequentially added and removed from the reaction.

Light is produced only when one of the nucleotide solution complements the first unpaired base of the template.

Sequencing by Synthesis (cont)

Limitation of the method is that the lengths of individual reads of DNA sequence are in the neighborhood of 300-500 nucleotides, shorter than the 800-1000 obtainable with chain termination methods (e.g. Sanger sequencing).

This can make the process of genome assembly more difficult, particularly for sequences containing a large amount of repetitive DNA.

Pyrosequencing AB (later Biotage, Qiagen) commercialized the process.

GS FLX, the latest pyrosequencing platform by 454 Life Sciences can generate 400 million nucleotide data in a 10 hour run.

History of Protein Sequencing

1934 Bergman degradation

late 1940s Edman degradation reaction (used for many years as the predominant method)

1958 Insulin sequencing by an enzymatic digestion process (Sanger, his first Nobel prize)

1989 Protein sequence by tandem mass spectrometer (MS/MS)

Electrospray ionization, ESI (Fenn, Nobel prize in 2002)

Matrix-assisted laser desorption/ionization.

Determining Amino Acid Composition

It is often desirable to know just the *unordered* amino acid composition of a protein before trying to determine the actual sequence.

This knowledge can be used to help the discovery of errors in the sequencing process or to distinguish between ambiguous results.

Steps of process:

- Hydrolysis – break up protein into its constituent amino acids (applying heat of 100-110 C° for 24+ hours)
- Separation – get the amino acid components

Protein Sequencing by MS/MS

Steps of process:

- Break the protein into peptides (using proteases, e.g. trypsin).
- Break down the peptides into fragment ions in a Tandem Mass Spectrometer (MS/MS).
- The mass spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones.
- Thus the spectrometer measures mass/charge ratio of an ion, and produces a spectrum.
- This spectrum is then used by a computer program attempting to determine the amino acid sequence.

Processing Spectrum of MS/MS

There are two major approaches:

- De novo peptide sequencing

  This is performed without using prior knowledge of any amino acid sequence. It is the process of assigning amino acids from peptide fragment masses of the protein.

- Database search

  This a protein identification process which uses prior knowledge of amino acids stored in a database, and attempts to find similarity between the spectrum provided by MS/MS and the spectrum of the proteins in the database.

De Novo Peptide Sequencing

Constructing a directed spectrum graph for a spectrum produced by MS/MS:

- Vertices:

  Since a mass s in an MS/MS spectrum was created by one of the ion types from $\Delta=\{\delta_1, \delta_2,\ldots, \delta_k\}$, a set of potential masses of the original partial peptide and – correspondingly – a set of vertices are generated for each value s of the spectrum:

$$V(s) = \{s+\delta_1, s+\delta_2, \ldots, s+\delta_k\}$$

De Novo Peptide Sequencing (cont)

- Vertices (continued):

  Therefore the complete of set of vertices for the spectrum graph:

$$\{initial\ vertex\} \cup V(s_1) \cup V(s_2) \cup \ldots \cup V(s_m) \cup \{terminal\ vertex\}$$

- Edges:

  For each vertex pair with mass difference of amino acid A, a directed edge (from smaller to larger mass) labeled w/ A is drawn.

Using the Spectrum Graph

The task is to find paths from initial vertex to terminal vertex.

There could be multiple such paths in the labeled spectrum graph.

Each path represents an amino acid sequence (read out from the labels).

A probability that peptide P represented by a sequence would produce the received spectrum S can be computed as

$$p(P,S) = \pi_{s \epsilon S} p(P, s)$$

where $p(P, s)$ is the probability of peak s .

The peptide with the highest probability can be chosen as the most likely sequence.

Pros and Cons of De Novo Sequencing

Advantages:

- Gets the sequences that are not necessarily in the database.
- An additional similarity search step using these sequences may identify the related proteins in the database.
- It is the best method for database search results validation. False positives are virtually eliminated this way.

Disadvantages:

- Requires higher quality data.
- Often contains errors.

De Novo Sequencer Implementations

Lutefisk

Johnson & Taylor, 1997; 19% peptide accuracy

http://www.hairyfatguy.com/lutefisk/

SHERENGA

Dancik et. al., 1999; 29% peptide accuracy

Peaks

Ma et. al. 2003; 25% peptide accuracy

www.bioinformaticssolutions.com/products/peaks/index.php

PepNovo

Frank & Pevzner, 2005; 30% peptide accuracy

http://proteomics.ucsd.edu/Software/PepNovo.html

Database Search

Steps of the algorithm:

A. Evaluates protein sequences from a database to compile the list of peptides that could result from each protein.

B. Determines the set of candidate peptide sequences that could meaningfully be compared to the spectrum by including only those which are near the mass of the observed peptide ion.

C. Projects a theoretical tandem mass spectrum for each candidate peptide.

Database Search

Steps of the algorithm (continued):

D. Compares these theoretical spectra to the observed tandem mass spectrum by the use of cross correlation (a measure of similarity of two waveforms, here the two spectra).

E. The candidate sequence with the best matching theoretical tandem mass spectrum is reported as the best identification for this spectrum.

Note: The algorithm works real well, if the protein did not go through multiple posttranslational modifications.

Post-Translational Modifications

Proteins – while involved in metabolic regulation – are subject to a large number of modifications.

Almost all protein sequences are post-translationally modified and about 200 types of modifications of amino acid residues are known.

A peptide fragment of a multiple times post-translationally modified protein produces a significantly different spectrum and therefore the above described identification algorithm will not find match with the spectrum of the original (unmodified) peptide derived from the database.

Virtual Database Search

Possible modification of the original algorithm:

In step B. not only determines the base-line set of candidate peptide sequences, but it also generates candidate peptides from all different possible multi posttranslationally modified version of the proteins.

The rest of the steps are the same.

Note: Leads to an unmanageable large combinatorial problem.

Another Approach

- Another possible modification of the original algorithm to handle the multi posttranslational modifications:
  - As an additional input to the algorithm the maximum number of allowed posttranslational modifications is also specified.
  - Instead of generating peptides from all possible modified proteins, it generates them only from the base proteins in the database (same as step B. in the original search algorithm).
  - Then in step D. (when comparing spectra) considers adjustments on the theoretical spectra (using dynamic programming).

Database Search Implementations

SEQUEST

Yates & Eng, 1994; it is a complete system and one of the first database search programs

http://fields.scripps.edu/?q=content/software

Mascot

Pappin & Perkins, 1999; it is a software search engine that uses MS data to identify protein from primary sequence databases

http://www.matrixscience.com/

Peaks

Ma et. al. 2003; it is a system that has also a database search software

www.bioinformaticssolutions.com/products/peaks/index.php

Database Search Implementations

X! Tandem

GPM, 2009; an open source software that can match tandem mass spectra with peptide sequences; simple-to-use, sophisticated application programming interface.

http://www.thegpm.org/tandem/

X!! Tandem

GPM, 2009; a parallel, high performance version of X!Tandem that has been parallelized via MPI to run on clusters or other non-shared memory multiprocessors .

http://wiki.thegpm.org/wiki/X!!Tandem