CSI5126. Algorithms in Bioinformatics
Projects — Fall 2018

Instructor: Marcel Turcotte

Version of October 30, 2018

1 Learning outcomes

- Critically assess a molecular biology problem and propose a bioinformatics solutions
- Further develop life-long learning skill
- Communicate technical information effectively, both in writing and orally

2 Deadlines

- 2018-10-29 — Project proposal (10%)
- Schedule to be determined — Project presentation (10 %)
- 2018-12-10 18:00 or earlier — Project report (30 %)

On Thursday, November 1, 2018, we will be discussing your project proposals. Follow the link below and put your name next to one or more of the available times to book an appointment with me.

- https://docs.google.com/document/d/1OEvgnRUQEdqCXclL4BmOdVV9A8gOSwjmIOZKyPiO-_E/edit?usp=sharing

3 Directives

3.1 Deliverable

The format of the projects is quite flexible. I foresee three broad types of work:

- The development of a novel application;
- The analysis of a new data set;
- A thorough review of the literature on a specific topic.

For all three types of work, I would like to see a review of the literature, sample data and a prototype implementation (where applicable). The main difference between each type of work will be the relative importance of each of the components.

3.2 Teamwork

Teams will be made of 1 or 2 members. Larger teams are possible and will have to produce proportionally more work! Complementary work between teams is also welcomed, i.e. two or more teams working on a related but complementary topic, leading to a more realistic application.

3.3 Report

The project is worth 50% of your final mark. Its marking will be based on the outline, a written report as well as a short presentation in class (10 minutes). Reports should be sufficiently detailed that it should be possible to implement the approach on the basis of the text alone. Having said that, you should also make every conceivable effort to keep the report concise. Assuming a team of size 2, a 10–15 page report should be appropriate. Suggested structure for the reports:

- Introduction
  - Background
  - Problem definition
  - Describing the data
- Methods
- Results
- Conclusions
- Future research
- Full list of references

3.4 List of Projects

Some of the projects require a fairly good background in statistics, I have annotated them with the letter S, while others may require more advanced knowledge of biology. Given the large number of projects there should be something for everyone. (A = application development, E = experiment, S = statistics, B = biology). You are also most welcomed to propose new projects.

1. **You own proposal**. By far the most interesting project for me is your own proposal. Hopefully, related to you thesis topic. This allows me to explore new areas of research.

2. **A deep learning-based approach to predict the localization of RNA molecules**

   RNA molecules are demonstrating a surprising breadth of biological functions. The repertoire of known non-protein-coding RNAs (ncRNAs) has grown extremely rapidly. Through all those discoveries, a new understanding of gene expression regulation is emerging.

   Herein, we take an important step to help better understand the cellular roles of RNA by predicting their sub-cellular localization. The advent of widely available frameworks for deep learning (e.g., Scikit-Learn and TensorFlow) as well the recent release of RNALocate, a database for RNA subcellular localization, are making this project possible and timely.

   Specifically, the project aims to answer the following research question: Can the subcellular localization of RNA molecules be predicted in silico from sequence information only?

   Artificial Intelligence has become a hot topic again. Through the groundbreaking work of Jeff Hinton (University of Toronto) and Yoshua Bengio (Universit de Montral), Canada is now playing a leading role in machine learning. This work has found applications in research, but also in the industry. So much so that all the leading high-tech companies (Google, Facebook, Microsoft, Thales, etc.) now have research laboratories in Canada.

   Throughout this project, the student will learn skills that are in high demand by the industry, locally, nationally and worldwide. He will develop skills preparing data for machine learning experiments. He will carefully design the research protocol to validate the results. He will evaluate the ability of different frameworks and deep neural network architectures to predict the localization of RNA molecules from sequence information only.

   The project will consist of the following steps: 1) Review the literature on deep learning in bioinformatics (2 weeks). Data preparation (3 weeks): Writing parsers to extract the sequence information from the RNALocate database. Prepare the data to remove redundancy. Design the cross-validation protocol (2 weeks). Carry out the experiment (3 weeks). Analyze the results (2 weeks): Writing the abstract. Creating the poster presenting this work.

- Angermueller, C., Prnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol Syst Biol* **12**, (2016).
- Jurtz, V. I. et al. An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* 33, 36853690 (2017).
- Min, S., Lee, B. & Yoon, S. Deep learning in bioinformatics. *Brief Bioinform* **18**, 851869 (2017).
- Zhang, T. et al. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res* **45**, D135D138 (2017).

3. **A deep learning-based approach to predict which RNA molecules will be packaged in exosomes**

   This project is similar in nature to the one above except for the data set. A machine learning approach will be used for the analysis of RNA sequences packaged in exosomes, which cell-derived vesicules.

   - Li, S., Li, Y., Chen, B., Zhao, J., Yu, S., Tang, Y., et al. (2018). exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Research* **46**(D1), D106D112. http://doi.org/10.1093/nar/gkx891.

4. **Using deep learning to predict sub-cellular localization of proteins**

   See Jurtz, V. I. et al. An introduction to deep learning on biological sequence data: examples and solutions. **Bioinformatics 33**, 36853690 (2017).

5. **Extensions for MotifGP**. We recently developed a tool for the inference of sequence motif (http://www.site.uottawa.ca/~turcotte/lab/software/motifgp/, https://github.com/mbelmadani/motifgp/). Here are some ideas that you could test:

   - The algorithm explores the space of regular expressions and identifies expressions that are enriched in the test set compared to the control set. At the end of the process, the regular expressions are converted into Position Specific Weight Matrices (PSWMs) in an *ad hoc* manner, the nucleotides are equiprobable. The project would consist of motifying the code that produces the PSWMs and use the actual matched sequences. The code is written in Python. It suffices to use matching groups in the expressions to collect sequence information and create PSWMs based on actual nucleotides frequencies. The hypothesis to be tested would that it would improve the E-value.
   - Currently, the positions of the motif are considered to be independant one from another. Recent research has shown that it could be advantageous to consider dependencies. Regular expressions in Python have *lookbehind* and *lookahead* extensions that could be used to implement dependencies. This could be done in various ways: including allowing the evolutionary algorithm to select where these dependencies should be added, or having depencies for all the positions. The project would consist of testing the following hypothesis: expressions with dependencies have better P-values.
   - MotifGP implements a multi-objective optimization. It would be interesting to explore adding new objectives: normalized complexity, entropy, GROMER, etc. See "C. K. Kibet and P. Machanick, Transcription factor motif quality assessment requires systematic comparative analysis, F1000Research, vol. 4, p. 1429, Mar. 2016." for ideas.

6. **Using deep learning to rank the structural motifs produced by Seed**

   Seed is a computer program that takes as input a set of unaligned RNA sequences and produces a set of secondary structure motifs. Suffix arrays are used enumerate complementary regions, possibly containing interior loops, as well for matching RNA secondary structure expressions.

   Seed has several criteria to rank the motifs that it produces: minimum description length, information theory, and free energy.

   In order to enable one of our future research directions, it would be interesting to see if deep learning can be used to rank the motifs produced by Seed.

   The project consists of: developping a deep neural network architecture, using information from Rfam to train the model, using a rigourous cross-validation strategy, compare the accuracy of deep learning to classify motifs against the existing criteria that used by Seed.

   - http://www.site.uottawa.ca/~turcotte/lab/software/seed/
   - https://github.com/turcotte/seed
   - http://rfam.xfam.org

7. **Identifying RNA-binding proteins (2)**. For this project, you will apply machine learning techniques to identify motifs that are specific to RNA-binding proteins. In would be nice if one of these tools was MotifGP (http://www.site.uottawa.ca/~turcotte/lab/software/motifgp/, https://github.com/mbelmadani/motifgp/). I can provide data sets if your would like to run experiments.

8. **Alignment-free sequence comparison**. Exact sequence alignment methods require $\mathcal{O}(n^2)$ time (and space) for comparing two sequences of length (n) each, which limits their application to relatively short segements. An alternative approach consists of comparing k-word frequencies of the two sequences. Implement and benchmark an alignment-free sequence comparison method. (Vinga et al. Comparative evaluation of word composition distances for the recognition of SCOP relationships. Bioinformatics (2004) vol. 20 (2) pp. 206-15; Sims et al. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc Natl Acad Sci USA (2009) vol. 106 (8) pp. 2677-82)

9. **Application of Deep Learning to identify RNA-binding proteins**

10. **Create an ML pipeline combining MotifGP and ModuleInducer**.

11. **Hardware acceleration of the median string alignment, for $k = 3$, where k is the number of input strings**. The median string algorithm can be used to align multiple sequences guided by a phylogenetic tree, amongst other things. As far as I know, there are no hardware accelerated version of this algorithm.

12. **Cloud computing.** Implement and benchmark a bioinformatics application in the cloud using Amazon Eastic Compute Cloud (or a competing technology). In the case of Amazon, public bioinformatics data sets are already available on Amazon's servers.

13. **MapReduce.** MapReduce is a programming model for implementing efficient data intensive algorithms on compute clusters. Implement and benchmark a bioinformatics application using MapReduce.

14. **CUDA.** CUDA is an architecture for developing general purpose programs for the NVIDIA graphics processing units (GPUs). The GPUs contain many cores, each capable of running thousands of threads, therefore offering a high level of parallelism at low cost. However, the architecture imposes many constraints — e.g. threads should be running in groups and follow the SIMD execution model. For this project, you should implement and benchmark a (simple) bioinformatics algorithm using CUDA. Currently, CUDA programs must be written in C.

15. **OpenCL.** OpenCL is an open and cross-plateform framework for developping parallel applications using GPUs, but also CPUs. Similarly to all three projects above, you should implement and benchmark a (simple) bioinformatics applications using OpenCL.
    www.khronos.org/opencl, www.macresearch.org/files/opencl

16. **Web Services in Life Sciences.** W3C defines a Web Service as follows: "a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically Web Services Description Language WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.". For this project, you can either review the literature on Web Services in bioinformatics and/or implement a client and/or implement a new Web Service.

    The BioCatalogue (www.biocatalogue.org) is a curated repertoire of Web services.

17. **Asymmetrical Substitution Matrices**. By construction, substitution matrices are created to be symmetrical. This makes sense since the matrices are mainly used for pairwise sequence alignments. Recently, multiple sequence alignment methods based on the (3) string median problem have been

developed that could perhaps make use of asymmetrical substitution matrices. The project consists of building asymmetrical substitution counts matrices from a sequence alignment dataset, such as BAliBASE, and quantify the asymmetry. In order to construct these matrices, one needs to reconstruct ancestral sequences, using the maximum parsimony principle, for example. Alternatively, the project could be more theoretical and look at the mathematical implications of this change (see Kosiol and Gojobori. Different versions of the dayhoff rate matrix. Mol Biol Evol (2005) vol. 22 (2) pp. 193-199).

18. **MSA with Asymmetrical Substitution Matrices.** This project would be the ideal complement to the one above and it consists of modifying an existing multiple sequence alignment method based on the string median problem so as to use an asymetrical substitution matrix, then benchmark the approch. The hypothesis to be tested is "using asymetrical substitution matrices improves the quality of multiple sequence alignment methods based on the string median problem.".

19. **Approximate Matching of RNA-RNA Secondary Structure Interaction Motifs** The project involves developing a pattern matching engine for a new class of motifs comprising multiple sequences. I have a prototype written in Java that can serve as a starting point.

20. **Grammatical Framework RNA-RNA Motifs**. Research on the usability of Simple Linear Tree Adjoining Grammars (SLTAGs) and Extended SLTAGs (ESLTAGs) for representing RNA-RNA motifs. What is the most economical grammatical framework that is sufficiently powerful to represent RNA-RNA interaction motifs? Are the parsing algorithms for these methods amenable to stochastic parsing?

21. **RNA-RNA Secondary Structure Motifs Editor.** The project involves developing a graphical user interface for interactively building RNA-RNA secondary structure motifs.

22. **Alternative Splicing**. Splicing is the process that removes the (introns) non-coding parts of a (pre-) messenger RNAs to produce the (mature) messenger RNAs, so that all the (exons) coding parts are juxtaposed contiguously. The mRNAs are then translated into proteins. For some protein-coding genes, there are more than one way to splice out the introns. Therefore, the same gene encodes more than one protein product. The rules that dictate which genes are alternatively spliced are not well understood. This project consists of studying properties of genes having a single variant vs genes having multiple transcripts. One such property could the number of (thermodynamically) stable stems near the boundary of introns and exons. Several data sets exist and can be used for the project; namely ASTD – Alternative Splicing and Transcript Diversity database, or EID www.meduohio.edu/bioinfo/eid.

23. **Speeding up the Simultaneous Alignment and Structure Prediction Problem**. Sankoff, Mathews, as well as our group have worked on the simultaneous alignment and structure problem. Unfortunately, the resulting algorithms are quite cpu intensive. Explore the possibility of accelerating these algorithms using bit-vector or shellable automata techniques.

24. **Alternative Translation Initiation Sites**. Wegrzyn et al. have proposed a method for the prediction lf alternative translation initiation sites, could structural motifs improve the prediction? (Wegrzyn et al. Bioinformatic analyses of mammalian 5'-UTR sequence properties of mRNAs predicts alternative translation initiation sites. BMC Bioinformatics (2008) vol. 9 pp. 232) Hypothesis: Covariance models are better predictors of alternative translation initiation sites than regular (linear) models? Methodology: A. Build a sequence data set consisting of the 5' UTR sequences of genes that have known alternative splice sites; see above paper for instance. B. Run a secondary structure discovery algorithm, and see if the sequences are enriched with secondary structure elements. An integrated model combining sub-models for all known translation initiation methods (Kosak, IRES, ribosome shunting...). See also: Rabani et al. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. Proc Natl Acad Sci USA (2008) vol. 105 (39) pp. 14885-90.

25. Towards an evolutionary-based score for the multiple sequence alignment problem. Using one of the existing benchmarks, such as BAliBase, HOMSTRAD, OXBENCH, PREFAB or SABmark, compare the following two strategies for solving the multiple sequence alignment problem. 1) The basis for the comparison will be the progressive sequence alignment method presented in class, and for which a Java implementation was given. 2) The new approach involves reconstructing the ancestral sequences at the

internal nodes of the tree, and performing a pairwise alignment of reconstructed ancestral sequences at the internal nodes of the tree. B. G. Hall. (2006) Simple and accurate estimation of ancestral protein sequences. *Proc Natl Acad Sci USA*, 103(14):5431–6, Apr 2006.

26. **Simultaneous Alignment and Structure Prediction of Circular RNAs**. Profile-Dynalign is a software system for the simultaneous alignment and structure prediction of two RNA sequence profiles. In its currentl implementation, this software system does not handle circular sequences. The project involves 1) a review of the literature for understanding how circular sequences are handled by the single-sequence approaches, such as mfold, 2) outline the necessary changes to Profile-Dynalign for handling circular sequences, and if time allows, implement and validate the change. (A)

27. **Synchronizing XML files: the bioinformatics way.** This project stems from two current trends. First, the documents of any individual are now squattered amongst several devices (work workstation(s) and server(s), home computer(s), PDA, iPod, AppleTV, GMail, MySpace, etc.). Second, XML documents are becoming ubiquitous (there are standards for office documents, on OS X, most configuration and application files are XML documents, web documents are XML documents, etc.). Synchronization at the document level results in a large number of conflicts, and offer little help to user to resolve the conflicts. XML documents can be represented as parse trees. In bioinformatics, several publications present framework to compare trees (edit distance on trees). This project consist of evaluating the feasibility of applying using the edit distance on trees as a way to synchronize documents. Ideally, a prototype (Java) implementation would be produced. Related ideas include saving the edit transcript on trees to save the history of the versions of a given document.

28. **Bit-vector algorithms**. The project consists of reviewing the literature on bit-vector algorithms for the approximate string matching problems, ideally developing a prototype implementation as well to illustrate the principles behind the technique. (Exploring the application of this technique to the RNA secondary structure problem would a plus).

29. **Vaccine bioinformatics**. Vaccines are widely used to protect populations from several diseases. The development of effective vaccines has therefore huge societal and economical impacts. Bioinformatics approaches are used to identify vaccine candidates. Most of the time, this means identifying, within pathogen genomes, short protein segments, 8–11 amino acids long, that have antigenic properties. The project consist of 1) a short review of the literature on bioinformatics approaches to vaccine discovery, or vaccine efficacy prediction. 2) Obtaining a dataset of peptides (short amino acids strings) with their antigenic properties. Developing and testing an approach for predicting antigenicity; this could be a consensus approach for instance. Here are two publications to help you start the project:

    - M. N. Davies and D. R. Flower. (2007) Harnessing bioinformatics to discover new vaccines. *Drug Discov Today*, 12(9-10):389–95.
    - J. A. Greenbaum, et al (2007) Towards a consensus on datasets and evaluation metrics for developing b-cell epitope prediction tools. *J Mol Recognit*, 20(2):75–82.

30. RNA Secondary Structure Plug-in for Bioclipse. Bioclipse is a Java application based on the Eclipse Rich Client Platform (RCP). As such, plug-ins can be written for extending this application. The project consists of developing a plug-in for drawing RNA secondary structure diagrams. I already have a Java implementation of the algorithm for drawing the secondary structure.

31. **Suffix arrays and inverted repeat finder (IRF - Gary Benson)**. Writing a new algorithm using Seed's library for finding inverted repeats. (A)

32. **Dynamic programming/suffix tree hybrid algorithms**. When the number of differences is bounded, hybrid algorithms, based on suffix tree and dynamic programming, can be built for finding the optimal alignment in $\mathcal{O}(km)$ time/space, where m is the size of the longest of the two input strings. This project requires developing a Java implementation of such algorithm using Daniela's suffix tree library, see Gusfield p. 268 for a description. (A)

33. **Take the ILP Challenge**. The following web site, www.protein-logic.com, has a series of datasets for three types of problems, functional class annotation, homology data, secondary structure prediction data. Design a novel approach for solving these challenges, which you will validate using the above mentioned data.

34. **Predicting Cellular Localisation**. Eukaryotic cells contain several sub-compartments, the "Cellular Localization" problem consists of predicting which compartment a protein is most likely to be found, on the basis of sequence information alone. The project may consist of a review of the literature and/or a novel analysis. LOCATE database, locate.imb.uq.edu.au, could be used to developing a new program. (E,B)

35. **Dangerous substrings.** This project exploits two related ideas that have been published recently. First, Greg Hampikian, in an interview for *Genome Technology* was saying: "He believes that there are some sequences that may be totally impatible with any life form and that such 'dangerous' sequences could play a big part in looking for potential drug targets. (...) If you have sequences that are intolerable for some pathogenic bacteria but that are common in humans because were are so far evolutionarily apart, then those can be used as drugs against the bacterial." Then, the following publication: S. Angelov, B. Harb, S. Kannan, S. Khanna, and J. Kim. (2007) Efficient enumeration of phylogenetically informative substrings. *J Comput Biol*, 14(6):701–23, defines a sequence tag as substring that is present in all the organisms of a subtree of a given phylogenetic tree but not in the other. Using the suffix tree library presented in class, write a (prototype) program to identify protein sequence tags.

36. **Consensus of consensus structures**. XDynalign is a software system for the simultaneous alignment and structure prediction of three RNA sequences. Sometimes, the number of available sequences is larger than tree. Under those circumstances, we would like to apply XDynalign to all the possible triples and build a consensus prediction. This project consist of implementing algorithms for extracting consensus statistics out of a set of consensus predictions (base pair frequencies, for instance), as well as applying the technique on two datasets (tRNA and 5S). (A)

37. **Zip codes** are referring to specific sequence/structure motifs found in the 3' UTR (untranslated region) of certain transcripts that are responsible for the routing of the transcript to a certain location of the cell. The concept is not well established, it re-occurs periodicly in the literature. The project consists of building a dataset of genes from five species of *Saccharomyces*, identifying a number of genes that are translated by polysomes at the vicinity of the mitochondria, and apply RNA secondary structure detection algorithms to hunt for conserved motifs. (E,B)

38. **Metabolic Pathways**. Proteins interact together to perform specific functions. Such network of interaction is called a molecular pathway. There are two main aspects to this field: how to infer/determine the connections and how to simulate cellular processes. There exists several computational approaches to model molecular pathways, including Petri-net.

39. **Regulatory-motifs**. Review of the literature on algorithms to automatically determine regulatory motifs (short sequence signals) in DNA sequence data. We have a Java library that can be used to implement a prototype application.

40. **SNP (Single Nucleotide Polymorphism)**. Review the literature of the methods for detecting SNPs, as well as their application. "Single nucleotide polymorphisms (SNPs) are common DNA sequence variations among individuals. They promise to significantly advance our ability to understand and treat human disease." (Excerpt from snp.cshl.org). See also Linkage analysis. (S)

41. **Molecular μ-arrays**. Today's technology (which borrows from inkjet technology) allows to fix tens of thousands of different macromolecules (DNA or protein molecules) onto a small surface. This technology allows to reveal which macromolecule is expressed, at different times, within different tissues, or different cellular states (disease vs non-disease). In the case of DNA chips, they measure the levels of expression of each gene.

42. **Mass spectrometry (MS)**. MS produces a spectrum of all the masses of all the compounds that are present in a sample. When an input protein is cut at specific sites, it will produce a specific spectrum. Such technology can now be used to fingerprint the content of a cell.

43. **Expression data + motif discovery**. DNA-μ-arrays allows to find genes that are simultaneously expressed. Those genes are most likely co-regulated, i.e. they share a common sequence signal in their promoter region. Daniela Cerna implemented a suffix tree library in Java, in the context of her honours project. Here, we would be re-using the library to help finding conserved motifs.

44. **Ontologies**. What is an ontology? What tools and knowledge representation formalisms (languages) are available to support the development of ontologies. Give examples of ontologies. Expose the problems associated with ontologies. An ontology is a controlled vocabulary (e.g. gene ontology). It allows to resolve some of the problems associated with data integration.

45. **Grammatical frameworks for RNA structure**. RNA secondary structure information can be represented using context-free grammars. As with most biological data, the information is better represented within a statistical framework. A “Stochastic Context-Free Grammar” (SCFG) has probabilities attached to its production rules. The two main issues with SCFGs are the parsing and the induction of the grammar. Review the literature on SCFGs (this includes COVE, infernal and pfold), and build a prototype parser in Java.

46. **Predicting Gene-Gene (Protein-Protein) interactions**. There exists a vast number of algorithms that allow to predict if two genes will be interacting. This includes: text-mining, co-location along the chromosomes, phylogenetic footprinting, etc.

47. **Lattice models**. Predicting the three-dimensional structure of a protein is a notoriously difficult problem. So much that alternative problems have been devised to circumvent it: secondary structure prediction, inverse folding problem, etc. Some authors have also been studying simpler systems, such as 2D and 3D lattices. Create your own implementation, this includes an algorithm to efficiently search the folding space and a scoring function. Run some simulations.

48. **Structure comparison methods**. Review the literature on 3D structure comparison. Implement at least one algorithm. Input: 2 three-dimensional structures, output: a measure of distance (typically root-mean-square deviation expressed in Å), and a list of equivalent residues.

49. **Bio-Ethics**. Bioinformatics deals with biological and medical data, according there are numerous related ethical issues: should patenting genes be allowed? How to handle patient data? How to deal with genomic data, imagine that the analysis of a dataset allows to draw conclusions about a population, a religious group, people who live in a specific region, etc. The consequences can be sever: it could be that this group will be more likely to suffer from certain diseases, such information could be used by insurance companies, employers, etc. to screen candidate.

50. **Survey and critical evaluation of the concept of open source code in bioinformatics**.

51. **Genome motifs viewer**. Construct a flexible graphical using interface to visualize shared motifs. Suggestions: make it 3D to ease viewing multiple strings. Motifs would be extracted from a suffix tree.

52. **Teaching tools**: interactive linear time construction of a suffix tree, showing the suffix links, interactive tools for software alignments.

53. **BioConductor**: is a bioinformatics environment developed using R. Evaluate the possible roles of this tool for teaching bioinformatics.

54. **Expectation-Maximization** (EM) algorithm and some of its applications in molecular biology. EM is used for training certain Hidden Markov Models, Covariance Models and building phylogenetic trees. What is it? What are the main applications? Prototype implementation. (S)

55. **Gibbs sampling**. This technique forms the basis for several motif detection tools. What is it? What are the main applications? Prototype implementation. (S)

56. **Bayesian networks**. What are bayesian networks? What is interesting about them? What are the bioinformatics applications of bayesian networks? Carry out a small experiment. (S)

57. **Predicting Phenotype from Patterns of Annotation, μ-arrays, etc.** One of the goals of bioinformatics research is to transform molecular biology into a predictive science. For example, given a certain pattern of gene expression, detected by μ-arrays for example, what would be the best treatment (personalized medicine)? Survey the literature on the use of bioinformatics techniques to assist medical diagnosis, prognosis and treatments. Where are we heading? When will personalized medicine be true? How much data? Remaining problems to be solved?

58. **Statistics behind BLAST**. Good candidate for a multiple teams work, where one team would focus on the statistics of word matching while the other would focus on hashing. Produce a Java implementation of hashing techniques for speeding up the sequence alignment problem. The part on the statistical analysis of hits requires a statistical background (S) but not the algorithmic part.

59. **Constructing phylogenetic trees**. For this project, you will produce a prototype implementation, in Java, of a modern method such as: quartet method, maximum likelihood or maximum parsimony. (s)

60. **RNA Secondary Structure and Phylogenies**. Explore the use of secondary structure information to improve phygenies inference.

61. **QSAR**. One of the main bioinformatics contribution to drug discovery is the Quantitative Structure Activity Relationship analysis (QSAR); the other is molecular docking. QSAR analyses take as input a set of compounds and their relative activity/efficacity. It then finds the commonalities between those molecules. The commonalities are then used to design new/better drugs.

62. **Genome assembly**. Because of physical limitations, only relatively short DNA sequences can be read (some 500 nt). For processing a complete genome, one approach, called shut-gun, consists of sampling small reads (500 nt) at random location along the chromosomes. The total number of reads is chosen so that the likelihood that each nucleotide is included into more than one read is high (typically each nt is part of 3, 5 or 10 reads). Computers are then used to stitch the reads together. One solution to this problem is related to the shortest super-string problem.

63. **Molecular docking** consists of predicting how two molecules will interact. This can either be two proteins or one protein and a small compound, such as a new drug. The two main factors that are taken into account are the shape and electrostatics of the two molecules.

64. **Tandem repeats**. Review the literature on tandem repeats detection and implement a prototype application. Tandem repeats are repeats of the form α^n, s.t. $2 \leq |\alpha| \leq 5$, in the case of micro-satellites, and each unit, α, is degenerated (which implies that the algorithms must allow for mismatches).

65. **Methods for detecting trans-membrane helices**. There is class of trans-menbrane proteins whose secondary structure can be reliably predicted. Those proteins are mainly made of α helices, such that if the loop connecting the helices i and $i + i$ is exposed to the inside of the cell, then the next one will be exposed to the outside of the cell. Use a Hidden Markov Model or Neural Network to reproduce this result.

66. **Surface/Interior**. Implement an algorithm to predict the solvent accessibility. Common choices for your implementation include: Neural Networks, Hidden Markov Models, and possibly decision trees.

67. **Protein Secondary Structure Prediction**. Implement a secondary structure prediction method and compare its accuracy to known methods. Common choices for your implementation include: Neural Networks, Hidden Markov Models, and possibly decision trees.

68. Comparing the accuracy and sensitivity of genetic programming algorithms vs your favorite machine learning approach (neural network, support vector machines) for **detecting parse segments** (coils) in protein sequences. (E)

69. **Can parsing improve the accuracy of protein secondary structure prediction methods?** There is evidence that the location of parse elements (coils) is more accurately predicted that helices and strands. Compare the accuracy accuracy your favorite machine learning algorithm (e.g. support vector machine or neural network) when trained on parsed vs unparsed segments. What are your conclusions/recommendations.

70. **Collecting and characterising RNA molecules with known secondary structure.** The secondary structure must have been determined experimentally or by co-variations studies. (B)

71. **Survey of the applications of cellular automata in bioinformatics**.

72. **Accurate Phylogenetic Reconstruction from Gene-Order Data**.

73. **Applications/benefits of the Semantic Web for the Life Sciences**.

74. **De Novo Identification of Repeat Families in Large Genomes** (See Pevzner ISMB 2005 paper, as well as www.drive5.com/piler).

75. **Survey of algorithms for the alignment RNA secondary structure**, possibly with prototype implementations.

76. **Modify the range of the constraints in X-Dynalign and evaluate the impact of the changes onto the accuracy and coverage**.

77. **Analysis of Hepatitis Delta Virus sequences using Profile-Dynalign**. Comparing the results with those derived by manual analysis of data in the laboratory of Dr Pelchat;

78. **Gene function prediction**. Design a new approach for predicting gene function from sequence. Test your approach against the system developed here: S. Ferré and R. D. King. (2006) Finding motifs in protein secondary structure for use in function prediction. *J Comput Biol*, 13(3):719–31.

79. β**-turn prediction**, predicting glycosylation sites, ...

80. **Kernel Methods** for ...

81. **Applications of conditional random fields** for ...

82. Other suggestions are most welcomed!

A Frequently Asked Questions (FAQ)

1. Can you suggest resources for finding information, including references, for our project?

   Consult the web sites of the major conferences in bioinformatics, many have their proceedings online.

   - Intelligent Systems for Molecular Biology [2009, 2008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000]
   - Research in Computational Molecular Biology [2009, 2008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000]
   - Pacific Symposium on Biocomputing [2009, 2008, 2007, 2006, 2005, 2004, 2003, 2002, 2001, 2000]
   - Past events

   Consult the sites of the major journals. Several journals will allow free access (based on your IP address, therefore you must use a UofO computer or a proxy account).

   - bio.site.uottawa.ca/wiki/space/Journals
   - Bioinformatics
   - Journal of Computational Biology
   - Others