

Edited Textbook: Bio-Molecular Computation

Edited by John H. Reif

Abstract

Biomolecular Computation(BMC) is computation done at the molecular scale, using biotechnology techniques. BMC is a new field, with largely unexplored methodologies. It is inter-disciplinary by nature, lying in the interface between biochemistry and computer science. We describe a number of distinct methods for doing BMC. All these methods use biotechnology techniques to do computation or processing at the molecular scale. This will be an edited text on BMC and related applications. The text will mostly describe RNA and DNA methods for BMC (this is also termed *DNA computation*).

Topics:

- the underlying biotechnology that BMC utilizes,
- a number of distinct paradigms for doing BMC,
- experimental techniques the field of BMC,
- applications of BMC, including related methods that do not necessarily involve computation,
- as well as theoretical results.

Goals of the Text. The goal is to instruct and inform the biology, chemistry and computer science communities about this emerging field.

Coauthoring by Small groups. The authors will be partitioned into very small groups of common expertise, and each group of authors will have a number of assigned sections. The members of each group will need to negotiate among themselves the co-authored writing of the assigned sections (Reif has some suggestions below), and keep Reif informed of their decisions and progress.

Format of Text. The text will typeset in latex, with contributed sections of chapters. The terminology will be defined consistently throughout the sections, and there will also be an appendix providing definitions of the terminology. Authors will be encouraged to liberally illustrate their sections. We will use latex to html conversion routines to provide also a web accessible version of the text providing addition illustrations, software, and lecture notes and other material. Each section will also provide references and problems.

Editing the Sections. In addition to editing by Reif, each section will be refereed and referee reports will be used to revise the sections. The authors will aid in refereeing the draft sections.

Schedule

1. One page Chapter Outlines and Abstracts of sections due July 1, 1999.
2. Detailed outlines due August 1, 1999.
3. Full length drafts of sections due Oct 15, 1999.
4. Referee reports of full length drafts due Nov 15, 1999.
5. Edited sections due Jan 15, 2000.

Organization.

- Chapter 1 will introduce BMC methods and briefly discuss the goals and potential applications of BMC by use of simple examples.
- In Chapter 2, we introduce biotechnology for BMC, including ways in which conventional biotechnology may need to be tailored for BM, as well as newly emerging biotechnology.

Then in successive Chapters we describe various distinct paradigms for BMC. We describe:

- in Chapter 3 the distributed molecular parallelism paradigm for BMC,
- in Chapter 4 the local assembly paradigm for BMC, and
- in Chapter 5 we describe alternative paradigms for BMC.
- Finally, In Chapter 6 we describe various applications of BMC

1 Introduction

1.0 Recombinant DNA (Author: Landweber). *We will provide a brief and very elementary introduction to Recombinant DNA.*

1.1 NP search problems. These are a class of computational problems apparently requiring a large combinatorial search for their solution, but requiring modest work to verify a correct solution. NP search problems may be solved by BMC by (i) assembling a large number of potential solutions to the search problem, where each potential solution is encoded on a distinct strand of DNA, and (ii) then performing recombinant DNA operations which separate out the correct solutions of the problem. *As an example, the text will briefly explain, in a simple way, the Adleman experiment on a Hamiltonian graph problem.*

1.2 Huge Memories (Author: Baum). BMC has the potential to provide huge memories. Each individual strand of DNA can encode binary information. A small volume can contain a vast number of molecules. We can perform massively parallel associative searches on these memories. *As an example, the text will provide an elementary introduction to methods for DNA data base search.*

1.3 Processing of Natural DNA (Author: the Princeton group). BMC techniques may also be used in problems that are not implicitly digital in nature, for example the processing of natural (biologically derived) DNA. These techniques may be used to provide improved methods for the sequencing and fingerprinting of natural DNA, and the solution of other biomedical problems. The results of processing natural DNA can be used to form *wet data bases* with re-coded DNA in solution, and BMC can be used to do fast searches and data base operations on these wet databases. *As an example, the text will provide an elementary introduction to re-coding of Natural DNA.*

1.4 DNA Cryptography (Author: the Duke and Mt Sinai groups). *The text will provide an elementary introduction to DNA Cryptography.*

1.5 Massively Parallel Computation (Author: Duke group). BMC also has the potential to supply massive computational power. BMC can perform massively parallel computations by executing recombinant DNA operations that act on all the DNA molecules at the same time. These recombinant DNA operations may be performed to execute massively parallel memory read/write, logical operations and also further basic operations on words such as parallel arithmetic. *As an example, the text will provide an elementary introduction to DNA XOR computation.*

1.6 DNA Nano-fabrication and Self-assembly (Author: the Self Assembly groups: Winfree, NYU and Duke group). BMC techniques combined with DNA nano-fabrication techniques may allow for the self-assembly of DNA tiles into lattices in 2 and 3 dimensions and the construction of complex nano-structures that encode computations. *As an example, the text will provide an example of a simple a Self-assembly tiling.*

2 Biotechnology for BMC

2.1 Nucleic Acid Structure (Author: Landweber).

- Describe the chemistry and dynamics of nucleic acids and enzymes,
- Introduce basic chemical properties of DNA as a molecule consisting of a linear sequence of nucleotides,
- Describe structure of RNA and certain proteins.

2.2 Nucleic Acid Kinetics (Authors: Gifford, Winfree).

- Describe the annealing of large strands of single DNA into double DNA, and the formation of complex 3D structures with secondary structure.
- Introduce the thermodynamics of DNA hybridization and denature operations.
- Describe software simulation of the kinetic models.
- Also discussions of dynamics and thermodynamics of RNA and certain enzymes.

2.3 Solution-based Recombinant DNA Technology (Authors: the USC group and Landweber). Biotechnology has developed a large set of procedures for modifying DNA, known collectively known as *recombinant DNA*. The most pervasive enabling biotechnology for BMC is solution-based recombinant DNA, that is the recombinant DNA operations are done on test tubes with DNA in solution.

- Basic principles of Solution-based recombinant DNA technology will be described and illustrated by graphic examples.
- Many recombinant DNA operations use hybridization and are specific to a DNA segment with a prescribed n-mer subsequence. We will describe
 - recombinant DNA operations including *cleavage* of DNA strands, *separation* of DNA strands, *detection* of DNA strands, and *fluorescent tagging* of specific DNA words.
 - operations that are not specific, including *ligation* of DNA segments to form covalent bonds that join the DNA strands together, *merging* of test tube contents, the denature operation, and separation by molecular weight.
- We will describe also the automation of Solution-based Recombinant DNA.
- We will discuss physical constraints for BMC using solution-based recombinant DNA including the volume, time, and energy.

2.4 Solid support and surface-based chemistry (Authors: the Wisconsin group). An example of an alternative recombinant DNA methodology is the *solid support* of individual DNA, for example by *surface attachments*. In solid support, the DNA strands are affixed to supports of some sort. In surface-based chemistry, surface attachments are used to affix DNA strands to organic compounds on the surface of a container.

- We will describe the automation of surface-based chemistry.
- We will also describe surface attachment methods that can be used for optical read-out (e.g., via fluorescent tagging of specific DNA words) on 2D DNA chip arrays.
- A possible drawback of surface attachment technology, in comparison to solution-based recombinant DNA techniques, is a reduction on the total number of DNA strands that can be used. We will discuss physical constraints for BMC using surface-based chemistry including the volume, time, and energy.

Enhancement of Recombinant DNA for BMC. BMC has certain requirements not met by conventional recombinant DNA technology. We will describe how to:

- modify conventional recombinant DNA— to obtain high yields and to allow for repeatability of operations.
- provide analytic and simulation models of key recombinant DNA operations.

2.5 Efficient Error-resistant Separations and Exquisite Detection (Authors: the USC group). Separation operations involve the isolation of all DNA with particular n-mer subsequences. Certain BMC methods require separation operations with high efficiency and high specificity.

- Approaches to solve this problem include the use of solid support, and the careful design of the n-mers used in separations may provide low error rates.
- Also, we will describe related methods for exquisite detection of molecules in low solution concentration.

2.5 Word Design for BMC (Authors: Condon, Winfree, Wood). is the problem of designing of a library of short n-mer sequences (DNA words) for information storage. Word design is crucial to error control in BMC. We will describe:

- the goals of good word design that minimize unwanted secondary structure, and minimize mismatching, by maximizing binding specificity,
- the conflicting requirements on word design for BMC: as strand length decreases (which is desirable), the Hamming distance between distinct words of information decreases (which is not desirable),
- various approaches for word design such as randomization, combinatorial search, and evolutionary biology techniques,
- software tools for word design.

3 The Distributed Molecular Parallelism Paradigm

We will introduce the *distributed molecular parallelism* paradigm for BMC, where the operations are executed in parallel on a large number of distinct molecules in a distributed manner, using the massive parallelism inherent in BMC.

Models for Distributed Molecular Parallelism (Authors: Adleman and Lipton).

- We will describe abstract models of molecular computation. The elements of *test tubes* are strings as in the case of DNA.
- These models allow a number of recombinant DNA operations on test tubes.
- We will describe how these models can be used as a simple programming language for BMC algorithms, and can execute circuit evaluation.

3.1 Solving NP Search Problems (Authors: Adleman, Lipton and Rochester group). We will :

- give a full description of Adleman's experiment, as well as subsequent improved techniques.
- describe Adleman's experiment in the context of these abstract models.
- describe other uses of the distributed molecular parallelism paradigm to solving NP Search problems:
 - finding satisfying inputs to a Boolean expression,
 - graph coloring problems,
 - integer factorization problem,
 - breaking the DES cryptosystem,
 - protein conformation.
- how the number of steps grows as a polynomial function of the size of the input, but the volume grows exponentially with the input.
- describe also the use of sophisticated heuristics which may result in a smaller search space and volume. (Ogihara and Ray).

3.2 Combinatorial Chemistry as NP Searches (Author: Landweber). *Combinatorial chemistry* techniques (also known as *diversity* techniques) have been used by biochemists to do combinatorial searches for biological objects with special properties. These techniques were very similar to the use of massive parallelism in BMC to solve NP search problems. Generally, they use recombinant DNA techniques to first construct a large pool of random sequences and then choose elements with specific properties from within the pool. We will describe:

- the techniques and applications of Combinatorial Chemistry and its use in BMC such as for word design.
- how the search space of combinatorial chemistry might be decreased by sophisticated heuristics used in NP search methods.

3.4 Associative Memory (Author: Baum). We will describe:

- Baum's idea for parallel memory where DNA strands are used to store memory words, and provided a method for doing associative memory searches using complementary matching.
- how this idea for associative memory can be extended to allow us to execute operations in parallel, on the words.

BMC Machines. We will describe how BMC machines using molecular parallelism and providing large memories, are being constructed at Wisconsin and USC. We will describe and compare their techniques. In both projects,

- a large number of DNA strands are used, where each DNA strand stores multiple memory words, and their operations on words include the Boolean logic operations.
- Both these machines are capable of performing, in parallel, certain classes of simple operations on words within the DNA molecules used as memory.
- Both projects use error-resistant word designs.

3.5 Surface Based BMC Machines(Author: the Wisconsin group). We will describe:

- how the Wisconsin project is employing a surface to immobilize the DNA strands which correspond to the solution space of a NP search problem.
- Since they are all on the same surface, all DNA strands are operated in a Single Instruction Multiple Data (SIMD) fashion.
- Their operations on words are restricted to mark, unmark, and destroy operations, which suffice for certain NP search problems.
- the key challenge in their approach: to provide scaling to a sufficiently large number of DNA strands within the constraints of surface attachment technology.

3.6 Solution Based BMC Machines (Authors: the USC group). We will describe:

- how USC project is doing computation without formation and breaking of covalent bonds, using a combination of solution-based and solid support methods, which are used to improve the efficiency of the separation operations.
- an abstract model of their operations and show that it can execute circuit evaluation.

3.7 Parallel Arithmetic (Authors: Mt Sinai and Pen groups). We will describe:

- the capability of BMC to quickly execute basic operations, such as arithmetic and Boolean operations, that are executed in single steps by conventional machines,
- how these basic operations should be executable in massively parallel fashion (that is executed on multiple inputs in parallel), and to permit chaining of the output of these operations into the inputs to further operations,
- an abstract model of surface-based computation and show that the surface-based model can execute circuit evaluation.

3.8 Theoretical Results for Distributed Molecular Parallelism (Authors: the Rochester and Duke groups). We will describe:

- a theoretical model that has a parallel associative matching operation, which provides for the combination of all pairs of DNA strings with subsequences that have a complementary match at a specified location.

- This PA-Match operation is very similar to the data base join operation.
- the use of this parallel associative matching operation for sequential speed-ups of sequential computations.
- the use of this operation for accessing shared memory in massiveley parallel BMC machines and in parallel circuit evaluation.

4 The Local Assembly Paradigm

We will describe how *local parallelism* allows operations to be executed in parallel on a given molecule (in contrast to the parallelism where operations are executed in parallel on a large number of distinct molecules but execute sequentially within any given molecule).

Before we describe these local assembly techniques, we first describe

- DNA nano-assembly techniques, and
- some previously known tiling results, which provided the intellectual foundations for local assembly.

4.1 DNA Nano-Fabrication Techniques(Author: The NYU group). We will describe:

- the design and experimental tests of nano-fabricated in DNA.
- examples of fabrication of various DNA junctions, crossovers, and polyhedra.
- the use of DX for rigidity or dsDNA for partial rigidity. We will describe the use of solid-support to avoid interaction between constructed molecules.
- various techniques used to verify the fabrication, including the use of jells, reporter strands, and atomic form miscropy.

4.2 Known Tiling Results (Authors: Winfree and Duke group).

We will describe (*domino*) *tiling problems* defined by Wang and various subsequent undecidable and hardness results.

4.3 DNA Self Assembly (Authors: Winfree and Duke group). We will describe:

- Winfree's idea of to doing tiling constructions by application of DNA nano-fabrication techniques.
- models for the assembly of the tiles is due to this hybridization of pairs of matching sticky pads on the sides of the tiles,
- advantages of self assemblies which advance with no intervention by any controllers,
- methods to constrain the geometry of the assembly and to allow for the placement of input DNA strands on the boundaries of the tiling assembly,
- methods to improve the speed and likelihood of successful assembly,
- experimental tests:
 - validating the preferential pairing of matching DNA tiles over partially non-matching DNA tiles, and
 - the DNA self-assembly of a non-computational 2D tiling.
- various proposed methods for doing computation using DNA self-assembly.
- simplified assembly schemes including small depth assemblies and string tilings, and give example of how to dointeger arithmetic via such tilings.

5 Alternative paradigms for BMC

We will describe other alternative paradigms for BMC:

5.1 The Splicing Paradigm (Authors: the Binghamton group). We will describe:

- describe the *Splicing* paradigm for BMC which provides a theoretical model of enzymatic systems operating on DNA, and has its roots in formal language theory and give examples of splicing computations and the generation of various classes of sequence sets.
- an experimental test of splicing.

5.2 RNA-based Paradigm for BMC (Author: Landweber). We describe the use of RNA rather than DNA as the basis of the biotechnology.

5.3 Bacteria-based Paradigm for BMC (Author: Landweber and Mt Sinai group). We describe the use of microorganisms such as bacteria to do computation.

6 Further applications of BMC

The field of BMC has restricted its attention mostly to applications which are hard computational problems, e.g., NP search problems, but the volume grows very quickly with the size of the problem. We will describe some other potential killer applications.

6.1 Processing Natural DNA: The DNA²DNA Paradigm (Authors: the Princeton group). We will describe:

- the re-coding of natural DNA to sequences of encoded n-mers, which can be then operated in a purely digital manner.
- how once natural DNA is re-coded, we can assemble large *wet data bases* containing DNA that encodes data of biological interest, without the problem inherent in I/O to an electronic medium.
- BMC, with its huge memory capacity, has a considerable advantage over conventional technologies for storing such biological data bases.
- Once the wet data bases are assembled then general BMC methods may be used to speed up many other key applications in biology and medicine, such as fingerprinting and mutation detection and other fast associative searches in these wet data bases.

6.2 DNA Sequencing(Author: the Duke and Princeton groups). We will describe:

- another possible application: DNA sequencing by hybridization, which is quite different to the enzymatic sequencing techniques commonly used.
- Redundant re-coding of n-mers may be used to reduce errors due to incomplete hybridize.
- These redundant encodings would be constructed and attached to the n-mers using known BMC methods, yielding an encoded array of n-mers providing the DNA sequence information
- We will also describe a divide and conquer approach to DNA sequencing.

6.3 DNA Cryptography: (Authors: Duke and Mt. Sinai groups). We describe tow techniques for using DNA to encrypt messages:

- one-time pad methods
- Stenography.