Chapter 16
Memetic Algorithms in Bioinformatics

Regina Berretta, Carlos Cotta, and Pablo Moscato

16.1 Introduction

Bioinformatics is an exciting research field for memetic algorithms (MAs). Its core activity is the integration of techniques from Computer Science, Mathematics and Statistics to address challenging computational problems related with the analysis of large volumes of data. Due to its huge relevance as a means to understand biology in the 21st Century, this field has attracted the attention of many pioneers in MAs, including the authors of this chapter.

During the past two decades, the field of molecular biology and the new high-throughput technologies associated with it has spawned a number of interesting problems. These problems can, in many cases, be posed as optimization problems which are combinatorial, non-linear, and often have aspects of both. Some examples arise in the analysis of large scale genetic datasets (e.g. gene expression using microarrays, massive datasets of single nucleotide polymorphisms derived from genome-wide association studies, etc.).

The field of bioinformatics is characterized by a constant evolution in computational methods for clustering and feature selection, analysis of phylogenetic trees

Regina Berretta
Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, School of Electrical Engineering and Computer Science, The University of Newcastle, University Drive, Callaghan, NSW, 2308, Australia
e-mail: Regina.Berretta@newcastle.edu.au

Carlos Cotta
Escuela Técnica Superior de Ingenieria Informática, Universidad de Málaga, Campus de Teatinos, 29071 - Málaga, Spain
e-mail: ccottap@lcc.uma.es

Pablo Moscato
Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine and Hunter Medical Research Institute, School of Electrical Engineering and Computer Science, The University of Newcastle, University Drive, Callaghan, NSW, 2308, Australia
e-mail: Pablo.Moscato@newcastle.edu.au

F. Neri et al. (Eds.): Handbook of Memetic Algorithms, SCI 379, pp. 261–271.
springerlink.com

(inference and reconstruction), image processing, protein analysis (structure prediction, sequence alignment), drug therapy design, among many others others. As we said before, many aspects of these problems are combinatorial in nature, involving the selection or the arrangement of discrete objects. Many of these combinatorial problems are NP-optimization problems, thus biologists are generally interested in finding the optimal solution of a given problem, but if that is impossible to obtain, they also rely for their investigations in high-quality solutions, provided by some metaheuristic technique. In this sense, MAs are a good strategy as they can provide solutions quickly, but then if they are coupled to an exact solver (thus forming a complete MA – check chapter 12), they can also prove the optimality of the final solution.

In general, researchers employ exact methods developed by themselves, and highly crafted for the problem at hand, or rely on Integer Programming reformulations of their problems. References in Mathematical Programming, Integer Programming for problems in computational biology can be found in works by Lancia [501] and Althaus *et al.* [15]. A hands-on approach to modeling using commercial packages can be found in [338] and [278]. Our experience with students, coming from different academic backgrounds, also suggest that the book by Williams [936], and the reviews of Greenberg, Hart and Lancia [332] and Festa [259], are not only useful but they have the added value of being very motivational for those interested in crossing fields and to jump into this new area. However, it is clear that since the size of the datasets associated to these challenges problems is, in general, is massive, in many cases it is necessary to develop efficient metaheuristics to deal with the large instances of these problems. As usual, research on metaheuristics is important as it can provide good upper bounding schemes to guide exact search procedures.

This chapter provides an review of MAs that have been developed to address some of the problems mentioned above. For an eagle's view of the contents, in Table 16.1 the reader can find a list of references grouped by application. For the sake of completeness we have also included in this table some applications in the wider area of biomedicine, where applications of memetic algorithms are also manifold. In particular, it is worth mentioning the deployment of MAs for optimizing cancer treatment, both in radiotherapy [347, 348] and chemotherapy [519, 520, 894]. Precisely related to this later issue of drug scheduling we can cite the work of Neri *et al.* for HIV multidrug therapy [658]. Imaging applications in tomography and imaging are also numerous [99, 144, 210, 211, 789] (please check [716] for a review of metaheuristic methods applied to microwave imaging). In the following sections we will focus on the purely bioinformatic tasks defined in the table though.

16.2 Microarray Data Analysis

With the introduction of DNA microarray technologies, it is now possible measure the expression of thousands of genes simultaneously. However, this obviously comes at a price as even a single microarray experiment leads to the need to deal with large datasets. This has posed a challenge primarily for statistics, as researchers

Table 16.1. An overview of MA applications in Bioinformatics

Area	Subarea	Reference
Microarray analysis	clustering	[406, 592, 698, 840, 841]
	gene ordering	[167, 576, 631]
	feature selection	[339, 402, 953, 964, 965, 966]
Phylogenetics	inference and reconstruction	[153, 155, 157, 298, 767, 937]
	consensus tree	[723]
Protein analysis	structure prediction	[53, 148, 150, 495, 496, 677, 790, 959]
	structure comparison	[107, 488]
Molecular design	ligand docking	[373, 612]
	PCR product primer design	[947]
Sequence analysis	DNA sequencing	[218]
	multiple sequence alignment	[883]
	supersequence problem	[151, 297]
Systems biology	cell models	[773]
	gene regulatory network	[465, 466, 671, 842, 843, 893]
Biomedicine	3D reconstruction of forensic objects	[789]
	Radiotherapy	[347, 348]
	Drug therapy design	[519, 520, 658, 894]
	Tomography	[99, 144, 210, 211]

now need to deal with the "large n, small m" problem (where n denotes the number of measurements on a single sample and m is the total number of samples). Statisticians obviously prefer to deal with the reverse situation, with more samples than measurements. When multi-variate methods are required, researchers resort to obtaining "molecular signatures", searching for a more coherent, reliable and robust set of molecular changes [668]. They count on Computer Science (allied of course with statistical methods) for the development of sophisticated algorithms to analyze such data.

The approaches for the analysis of microarray datasets can be primary classified as *unsupervised* and *supervised methods*. At this description level, we can understand that these microarray datasets are basically two-dimensional arrays of values (the measurements) and that a re-assignment of labels to the samples (and, analogously, to the measurements) helps to uncover some structure within the data.

Clustering algorithms are the most common example of unsupervised methods to find these structures. Another unsupervised method, which can be seen as a

particular type of clustering algorithm is called *gene ordering*. In this case the overall objective is to find a permutation of either the rows or columns of this two-dimensional array such that those having the same patterns of global expression are relatively close in the permutation. An example of supervised method is feature selection, in which the aim is selecting a subset of features (genes in this case) such that a main goal is optimized, for example, classification accuracy.

We now give a brief description of some MAs that have been proposed to address the clustering and feature selection problems in microarrays.

16.2.1 Clustering

From the description we have given before, it is clear that clustering encompasses a wide number of different problems, as the word "scheduling" in Production Planning and Operations Research encompasses different specific problems. Merz and Zell's proposal [592] for the clustering problem in microarray data analysis is based on a model in which the task is to define an assignment of objects into clusters, such that the sum of squared distances to the centroid of the cluster is minimized. They proposed a MA which uses the K-Means algorithm as a local search technique. They use uniform crossover and they also propose a new one denominated replacement recombination operator. They compare the MA with a multi-start k-means local search using five different microarray datasets.

Speer *et al.* used in [840, 841] a Minimum Spanning Tree (MST) to represent the data, where each node is a gene and each edge between nodes i and j represent the dissimilarity between genes i and j, thus modeling the clustering problem as tree partitioning problem, i.e., deleting a set of edges to find the clusters. They proposed a MA based on the framework presented by Merz and Zell in [592]. They use two fitness functions, the sum-of-squared-error criteria (the same used in [592]) and the Davies-Bouldin-Index [186], which minimizes the intra-cluster and maximizes the inter-cluster distances. Using four microarray datasets, they compared the MA with two other popular clustering algorithms, the average linkage algorithm [242] and the Best2Partition [950], which is also based on a MST-representation of the data.

Palacios *et al.* [698] present the results of different population based metaheuristics (genetic algorithms, MAs and estimation of distribution algorithms) to obtain biclusters from microarray datasets. According to the authors, the advantage of finding biclusters in microarray datasets (instead of traditional clusters) stems from the ability to find a group of genes that are similar in a specific subset of samples. To analyze the performance of each algorithm, they used a yeast expression dataset comprising 17 samples on 2,900 probes.

Gene Ordering is another unsupervised method that can be interpreted as a special type of clustering algorithm. The objective is, given a gene expression dataset, to rearrange the genes, such that genes with similar expression patterns stay close to each other. MAs to tackle this problem have been proposed in [167, 576, 631]. In [167], Cotta *et al.* represent a solution as a binary tree, using hierarchical clustering as a start point. The crossover operator is similar to the one used in [155],

using subtrees from the parents to create an offspring. Flipping subtrees are used as the model for the mutation operator. Two local searches are applied, the first one works by inverting branches of subtrees and the second one employs a pairwise interchange local search. They test the MA in instances with up to 500 genes. Mendes *et al.* [576] uses the same MA, but with the objective to evaluate the impact of parallel processing in the performance of the MA and ability to apply it in larger instances (up to 1,000 genes). More recently, in [631] these MAs are improved significantly, with the inclusion of new local searches which employ Tabu Search. The MA is tested not only in microarray instances (containing more than 6,000 genes), but as well in images, where the objective is unscramble the rows of an image when the image has all its rows permuted at random. The images are excellent as benchmark instances and help to evaluate gene ordering and different clustering algorithms, making it easier to understand the quality of the results. The MA proposed by Moscato *et al.* [631] has been successfully applied in different microarray studies [63, 170, 330, 397, 577, 768].

16.2.2 Feature Selection

Feature selection methods are used primarily in bioinformatics to reduce the dimensionality of a dataset to help to discriminate between classes of samples under study. We note that the definition of a feature is rather general, it can be a gene expression (as in microarray datasets), a single nucleotide polymorphism (SNP) (as in genome-wide association studies), protein abundances (as in ELISA kit panels), among many others sources of biological information. Feature Selection methods can be classified as *filter* or *wrapper methods*. In filter methods, the features selected are evaluated based only on the characteristic of the data and in the wrapper methods, a classification algorithm is embedded in the method, giving constant feedback regarding the quality of the set of features selected.

Zhu *et al.* [965] present a MA for feature selection problems with the objective to improve classification performance. Each individual in the population is composed of a set of selected features (X) and a set of excluded features (Y). The local search procedure move features between sets X and Y based on some filter ranking methods, such as ReliefF, Gain Ratio and Chi-Square. They evaluated the performance of their approach using four UCI datasets (UC Irvine Machine Learning Repository[1]) and four microarray datasets, showing improvements in the classification accuracy.

In [953], Zhu and Ong present a similar MA, but now using a Markov blanket approach in the local search procedure. In [964], the same authors present a comparison study between the MAs presented in [965] and [953]. They evaluated the results on synthetic and real microarray datasets. Both MAs perform well in regards to classification accuracy, but the one that uses Markov blanket approach gives smaller feature sets. Finally, in [966], they present a memetic framework that combines the previous approaches with a hybridization of wrapper and filter feature selections methods. The computational tests were done in fourteen microarray

[1] `http://archive.ics.uci.edu/ml/`

data sets containing 1,000 to 24,481 genes. They have also tested their methods for hyperspectral imagery classification. The classification accuracy was good and the number of features selected varies depending on the local search used.

Other MAs for feature selection problems were proposed in [339, 402]. However, as stated by Zhu *et al.* [966], due to the inefficient local search methods a large amount of redundant computation is incurred on evaluating the fitness of feature subsets. This is an issue worth considering in detail when designing an MA as we rely on the power of local search, associated with good data structures, to speed-up the process. This is an area of great interest and we hope more sophisticated MAs will be developed during this decade.

16.3 Phylogenetics

The aim of plylogenetics is to study the evolutionary relationship between species, which can be represented by a phylogenetic tree. The inference of phylogenetic trees, known as Phylogeny Problem, is a very challenging task and is certainly important in molecular biology. It has connections with other problem domains in bioinformatics like *multiple sequence alignment*, *protein structure prediction*, among others [153]. The aim of the Phylogeny Problem is to find the tree (or in certain cases the network), that best represents the evolutionary history of a set of species. Several criteria have been defined in order to measure the quality of a certain tree given certain input data (typically, molecular data corresponding to a collection of different organisms or taxa); these can be broadly grouped into sequence-based methods (such as maximum parsimony and maximum likelihood) and distance-based methods (e.g., minimal ultrametric trees). Unfortunately, *NP*-hardness has been shown for phylogenetic inference under most of these models [190, 191, 277, 942]). Due to the complexity of the problem, the research focuses in the development of powerful metaheuristics, like MAs [153, 155, 157, 298, 767, 937].

Cotta and Moscato proposed several MAs for hierarchical clustering from distance matrices under a minimum-weight ultrametric tree model (i.e., finding an ultrametric tree of minimal overall weight, such that its associated distance matrix bounds the observed distances from above). The first approaches [155] were based on the use of evolutionary algorithms endowed with heuristic decoders, which could be viewed as greedy hill-climbers for genotype-to-phenotype mapping. Although these provided much better results than other simpler decoder-based approaches and tree-based evolutionary algorithms, their computational cost was also large. Later [157] an orthodox memetic approach was presented based on the use of a tree representation and a local search operator based on tree rotations.

A scatter search method using path relinking was subsequently presented by Cotta [153]. Scatter Search (SS) [314, 320, 500] is a powerful metaheuristic which can be considered as a particular type of MA that often relies more on deterministic strategies rather than randomization. In this work, the author used a ultrametric model and a minimum weight criterion as in previous works [155, 157]. The SS

algorithm was evaluated using five real biological data sets from an online repository –the TreeBase site[2]– and was shown to compare favorably to an evolutionary algorithm and a MA. Related to this, Gallardo *et al.* [298] propose an hybrid algorithm that combines Branch and Bound (BnB) and MA in an interleaved way. The idea is to have both techniques sharing information between them. They used the same five biological data sets from as [153] and showed improved results.

Williams and Smith [937] use maximum parsimony as the optimization criteria, which means that the tree with the least evolutionary events is the best. They propose a MA, which uses diverse and elitist populations (similar with the ones used in scatter search methods). More precisely, their approach is based on maintaining a collection of Rec-I-DCM3 trees (Recursive-Iterative DCM3, a powerful heuristic for designing maximum parsimony trees [777]) which cooperate within a selectorecombinative evolutionary algorithm. They evaluate their method using biological datasets with up to 4,114 sequences, obtaining better results than parsimony ratchet [669] and TNT (Tree Analysis using New Technology[3]). Richer *et al.* [767] also uses maximum parsimony as the optimization criteria. They propose a MA that uses progressive neighborhood as local search (similar with VNS - variable neighborhood search [364]). They used twelve instances from TreeBase, and obtained results that were generally equal or better than TNT.

A problem related to phylogenetic inference is that of finding consensus trees, namely finding a tree that summarizes the information comprised in a collection of trees (e.g., finding a unique tree that faithfully amalgamates the outcome of different phylogenetic inference methods). A seminal approach to this problem using evolutionary methods can be found in [152] on the basis of the TreeRank distance measure [916] between trees. Pirkwieser and Raidl [723] tackled this problem using VNS, evolutionary algorithms (EAs) , MAs (using EAs endowed with local search on different tree-based neighborhood structures), and multi-level hybrids based on the intertwined execution of VNS and EA/MA which ultimately produced the best results.

16.4 Protein Structure Analysis and Molecular Design

Problems involving analysis of protein structure are fundamental in bioinformatics. We refer to Oakley *et al.* [677] who present a review of problems involving analysis of protein structure (including structure prediction, structure comparison, aggregation of structures, etc.).

The protein structure prediction (PSP) problem aims to find the 3D structure with minimum energy (based in a specific energy model) given the primary sequence of the protein (i.e., the linear sequence of amino acids composing the protein). Krasnogor *et al.* [495] analyzed three main factors affecting the efficacy of evolutionary algorithms for PSP: the encoding scheme, the way illegal shapes are considered by the search, and the energy (fitness) function used. In [148] the protein structure

[2] `www.treebase.org`

[3] `http://www.zmuc.dk/public/phylogeny/tnt/`

prediction problem on the hydrophobic-polar (HP) model was considered. The HP model [213] is based on classifying each amino acid into two classes: hydrophobic or non-polar (H), and hydrophilic or polar (P), according to their interaction with water molecules. In this case the binary sequence of H/P amino acids is embedded in a cubic lattice subject to non-overlapping constraints, with the aim of maximizing the number of H-H contacts, namely the number of H-H pairs that are adjacent in the lattice. The MA featured the inclusion of a backtracking operator in order to repair infeasible protein configurations. A similar approach was used in [150] in the context of the HPNX energy model, an extension of the HP model in which polar amino acids are split into three classes: positively charged (P), negatively charged (N), and neutral (X). Krasnogor *et al.* [496] presented a multimemetic algorithm for protein structure prediction using four different models (HP in square and triangle lattice, and functional model proteins in the square and diamond lattice). Bazzoli and Tettamanzi [53] also considered the HP cubic lattice model. They presented a MA using a self-adaptive strategy, where the local search is applied with a probability guided by a function similar to the one used in simulated annealing, with the aim to either control exploitation or diversification. According with the authors, the MA was strongly based on the MA proposed by Krasnogor and Smith [491], where the authors compared self-adaptation against other local-search approaches for the traveling salesman problem. Santos and Santos [790] presents a MA for the protein structure problem using 2D triangular HP lattice model, whose main feature was the use of caching in order to reuse computation and speed-up fitness evaluation. The study of Zhao [959] addressed HP models as well. They described several metaheuristics such as MAs, tabu search, ant colony optimization, self-organizing map-based computing approaches and chain growth algorithm PERM, highlighting their advantages and disadvantages.

Protein structure comparison or protein alignment is another important problem in the area of protein structure analysis problem. In this case the goal is to identify structural similarities between proteins. Some MAs developed to deal with this problem can be found in [107, 488, 568, 911]. Carr *et al.* [107] considered the maximum contact map overlap problem. They presented a multimemetic algorithm where a family of local searches is used: selection of the particular local search to be applied depends on the instance, stage of the search or which individual is using it. The MA proposed is a combination of the genetic algorithm proposed by Lancia *et al.* [502] and six different local searches. Their computational results have showed that the results obtained by their method are compatible with the state of art in this problem. Also, Krasnogor [488] proposed a self-generating MA to obtain structural alignment between pair of proteins using the Maximum Contact Map Overlap (MaxCMO) problem as a model. MaxCMO is an alignment of two proteins that maximizes the structural similarity. They tested the approach in four different data sets, of which one was composed of randomly generated proteins and the other three data sets with real world proteins.

A bioinformatics area closely related to protein structure analysis is that of molecular design, which actually can be regarded as a superset of the former. Indeed, conformational analysis, namely determining the low-energy configurations

a molecule can adopt is a natural generalization of the protein structure prediction problem (for example, Zacharias *et al.* [954] presented a MA based on a genetic algorithm endowed with simulated annealing to determine the ground state geometry of molecular systems). In general, molecular design is a very hard problem, and numerous evolutionary approaches have been proposed in the literature to deal with problems in this area, e.g., [128, 935].

Ligand docking, i.e., the identification of putative ligands based on the geometry of the latter and that of a receptor site, is a problem within the area of molecular design with paramount interest for structure-based drug discovery. MA approaches to this problem have been proposed by Hart *et al.* [373, 612] using an evolutionary algorithm endowed with the Solis-Wets method for local search (see Chapter 12), aimed to minimize the free energy potential of the docking. This MA is used in the AutoDock[4] software package. MAs have also been used for PCR (Polymerase chain reaction) product primer design [947], taking into account constraints such as primer length, GC content, melting temperature, etc.

16.5 Sequence Analysis

Sequence analysis is arguably one of the lowest-level tasks in bioinformatics, albeit it remains a very important one due to its role in generating the input data for further biological problems. Within this general subarea we can cite problems such as DNA sequencing and the alignment of genomic/proteomic sequences.

DNA sequencing amounts to determining the correct order of nucleotides in a certain DNA sequencing. This order must be ascertained by assembling short fragments of DNA obtained from the fragmentation by chemical or mechanical means of a larger sequence. These fragments are typically randomly distributed across the sequence and partially overlap, thus leading to a permutational problem with strong similarities to that of finding a minimum weight Hamiltonian path. In [218] a spatially-structured evolutionary algorithm endowed with a so-called problem-aware local search (PALS) procedure is presented for this purpose.

Another important problem in sequence analysis is that of aligning sequences of nucleotides or amino acids. This problem actually bears some relationship with sequencing, since the determination of the best overlap among DNA fragments requires finding the best pairwise alignment. The applications of sequence alignment are not limited to this case though; thus, they are very important in phylogenetic studies to cite a relevant example. This alignment problem is easily solvable in polynomial time for two sequences using a dynamic programming approach, but its complexity quickly grows for when a multiple sequence alignment is sought. Not surprisingly, evolutionary methods have been commonly applied to this problem – see [813] for a survey. Some of these evolutionary approaches can be actually regarded as memetic. For example, the evolutionary Clustal/improver presented in [883] incorporates a seeding mechanism (using the outcome of the Clustal[5] software

[4] `http://autodock.scripps.edu/`

[5] `http://www.clustal.org/`

package) for creating a high quality initial population, and an improvement strategy based on the removal of matched gap columns which can be regarded as a simple form of local search.

Closely related to sequence alignment, the problem of finding the shortest common supersequence (SCS) for a collection of biological sequences stands as another important task. A supersequence of a given sequence is a possibly longer sequence in which all the symbols of the former can be found in the same order (although not necessarily consecutively). Finding the SCS for a given collection of sequences is a NP-hard problem that has been commonly dealt with in metaheuristics [70, 83, 149] including MAs. Thus, Cotta [151] considered a MA defined on the basis of an evolutionary algorithm endowed with a repairing mechanism (based on a greedy heuristic) and a local search operator based on the iterative removal of symbols in the tentative supersequence. Later, Gallardo *et al.* [297] presented a multi-level MA that combined the previous algorithm with a beam search algorithm (see Chapter 12), executed in an intertwined way. This MA was shown to provide much better results than the combined algorithm as stand-alone techniques.

16.6 Systems Biology

Systems biology [13] is a prominent interdisciplinary area of bioscientific research focusing on the holistic study of cellular systems from the perspective of (and using tools from) complex systems and dynamical systems theory. This encompasses the analysis and modeling of cell systems, including the study of networks of genomic/proteomic/metabolomic interactions. The latter are very amenable to the use of network-theoretical results and graph-based algorithmic tools, among which MAs excel. Thus, Spieth *et al.* consider a memetic approach to gene regulatory network modelling using linear weight matrices [924] and S-systems [914]. They use a binary genetic algorithm to evolve the topology of the network, and an evolution strategy to do local search on the parameters of the model representing the network. They consider a so-called *feedback MA* in which the outcome of the local search is used to filter gene dependencies whose strength is below a certain threshold. This can be regarded as a Lamarckian learning procedure, as opposed to the Baldwinian learning of the simpler MA [842] without feedback. An analogous approach is followed by Norman and Iba [671]: they consider time series data of gene expression and use a differential evolution endowed with hill climbing to determine the structure of the network and the kinetic parameters; an information-based criterion is used for fitness evaluation. It is also worth mentioning the work of Kimura *et al.* [465] in which a genetic local search method is used to solve the inference problem in the context of S-systems. In a later work [466], they consider a cooperative approach based on multiple subpopulations and problem decomposition and use golden section search in order to do local improvement. Tsai and Wang [893] consider a differential evolution hybridized with local search for S-system inference too.

A wider perspective on cell models is provided by [773]. They consider the use of P-systems [738], a computing model included in the ampler paradigm of membrane

computing [739]. These computational models are inspired by cellular processes, and can be roughly described a system of so-called *membrane structures*, namely permeable (and potentially nested) containers that comprise collections of symbols and grammar-like rules for their evolution. By an appropriate definition of the rules and a wise arrangement of membranes it is possible to carry out an arbitrary computation. The biological inspiration of these systems make them specifically suited for cell modelling and simulation though. Romero-Campero *et al.* use a two-level genetic algorithm to evolve the structure of a P-system: the upper level is devoted to searching in the space of rules, and the lower level performs numerical adjustment of the kinetic parameters determining the probability of application of each rule.

Acknowledgements. C. Cotta is supported by Spanish MICINN under project NEMESIS (TIN2008-05941) and Junta de Andalucía under project TIC-6083.