DNA Sequencing Hybridization Based on Multi-Castes Ant System

Cyrille Bertelle,
LIH, Le Havre University
BP 540 76058 Le Havre
cedex, France
Cyrille.Bertelle@univ-lehavre.fr

Antoine Dutot *,
LIH, Le Havre University
BP 540 76058 Le Havre
cedex, France
antoine.dutot@laposte.net

Frédéric Guinand
LIH, Le Havre University
BP 540 76058 Le Havre
cedex, France
Frederic.Guinand@univ-lehavre.fr

Damien Olivier
LIH, Le Havre University
BP 540 76058 Le Havre
cedex, France
Damien.Olivier@univ-lehavre.fr

ABSTRACT

The problem of DNA sequencing by hybridation methods is considered in this paper when all kinds of errors are taken into account. A distributed method based on dynamic organizations of reactive agents is used in an environment composed by a SBH-graph (Sequencing by Hybridation). Explorations of such graphs are studied with contraints representing the possible errors comming from the DNA sequencing operation. With these explorations, we finally find a set of sequences that we try to minimize and which have to containt the original sequence.

Keywords

DNA Sequencing by Hybridization, SBH-graph, Agent, MAS, Ant Algorithm, Dynamic Load-Balancing, Distributed Computing.

1. A GRAPH MODEL FOR DNA SEQUENCING

The sequencing of DNA is one of the basic operation in molecular biology. It consists in determining the sequence of bases (or nucleotides) of a given target piece of DNA. Among the existing methods designed for that purpose, we are interested in *sequencing by hybridization* (SBH) [12, 1, 6]. From the data produced by this method it was recently shown that we can build a new kind of graph called SBH-graph [10].

The associated computational problem consists in rebuilding the original sequence from this graph. Its consists to explore the SBH-graph with constraints. Unfortunately, restricted versions of the problem are NP-Hard [9, 2], and it is moreover not always possible to determine one solution for some instances of the original version of this problem. So, instead of trying to find the best or the most probable solution, we try to find as many potential solutions as possible.

1.1 Graph Formulation

The problem of rebuilding a sequence using a set of subsequences obtained by the *sequencing by hybridization* method can be formulated as a graph theory problem.

Many works have been done in the field and several graph formulations have been proposed [13, 9, 2]. One graph model able to take into account all the characteristics of the data produced by the SBH process is described in [10].

The model is a directed graph $G = (V, A)$ such that V is the set of vertices and A the set of edges. Two types of vertices are distinguished: type a and type b. In the remaining part of the paper, vertices of type a will be represented by single-circled nodes and vertices of type b will be represented by double-circled nodes in figures. Vertices of type a are gathered into the set V_a and vertices of type b are gathered into the set V_b. So, $V = V_a \cup V_b$. We denote $C_k = (v_1, v_2, \ldots, v_k)$ a path of length k in G. $V_a(C_k) = \{v_i \mid v_i \in C_k \text{ AND } v_i \in V_a\}$ (resp. $V_b(C_k) = \{v_i \mid v_i \in C_k \text{ AND } v_i \in V_b\}$) is the set of vertices of type a (resp. b) belonging to C_k. By extension, we define $k_a = |V_a(C_k)|$ (resp. $k_b = |V_b(C_k)|$) the cardinality of the set $V_a(C_k)$ (resp. $V_b(C_k)$). As $V_a(C_k)$ (resp. $V_b(C_k)$) is a set, any vertex v_i appears at most once within $V_a(C_k)$ (resp. $V_b(C_k)$). Then, given L, N_a, N_b and

*Supported by French Ministry of Research Grant

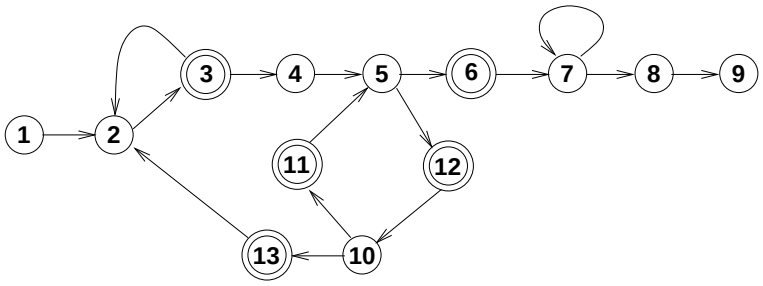

Figure 1: Example of a SBH-graph $G = (V, A)$ such that $V = V_a \cup V_b$ with $V_a = \{1, 2, 4, 5, 7, 8, 9, 10\}$ and $V_b = \{3, 6, 11, 12, 13\}$.

C_k	k	k_b	k_a	**is a solution**
$(1, 2, 3, 4, 5, 6, 7, 8, 9)$	9	2	7	No (too short)
$(1, 2, 3, 2, 3, 4, 5, 6, 7, 8, 9)$	11	2	7	Yes
$(1, 2, 3, 2, 3, 4, 5, 6, 7, 8)$	10	2	6	No (k_a too small)
$(1, 2, 3, 2, 3, 2, 3, 4, 5, 6, 7, 8, 9)$	13	2	7	No (too long)
$(1, 2, 3, 4, 5, 6, 7, 7, 8, 9)$	10	2	7	Yes
$(1, 2, 3, 4, 5, 6, 7, 7, 7, 8, 9)$	11	2	7	Yes

Table 1: Some paths in the graph

ΔL, ΔN_a and ΔN_b 6 positive values, solving the problem consists in finding all paths C_k in G such that:

$$L - \Delta L \leq k \leq L + \Delta L \tag{1}$$

$$N_a - \Delta N_a \leq |V_a| - k_a \leq N_a + \Delta N_a \tag{2}$$

$$N_b - \Delta N_b \leq k_b \leq N_b + \Delta N_b \tag{3}$$

REMARK 1. *all paths C_k start from the same fixed vertex.*

1.2 Example

We consider the example presented in Figure 1. The six parameters of the problem are $L = 11$ and $\Delta L = 1$, $N_a = 1$ and $\Delta N_a = 0$, $N_b = 2$ and $\Delta N_b = 2$.

We are looking for paths such that their lengths are equal to k with $10 \leq k \leq 12$. They must contain no more than 4 double-circled vertices, which can be expressed as $k_b \leq 4$ and they must contain $|V_a| - 1$ distinct nodes belonging to V_a which is equivalent to $k_a = 7$. In the table 1, characteristics of some paths in the graph are presented together with their validity with respect to the relations (1), (2) and (3). All paths begin with the same vertex (1) following Remark 1.

1.3 Graph Path Definitions

We define in this section different kind of paths which we are going to use in the remaining part of the paper.

DEFINITION 1. *a* symbolic repetition *in a path is a series of consecutive vertices which number of copies is a symbol. The symbol may be a '+' or a '*'. The '+' indicates a number of copies greater than or equal to 1, the symbolic repetition is said* effective*, while the '*' indicates a number of copies greater than or equal to 0.*

For instance: $(4\ 3), \emptyset, (4\ 3\ 4\ 3\ 4\ 3)$ are instances of the symbolic repetition $(4\ 3)^*$.

DEFINITION 2. *a* symbolic path *is a path containing symbolic repetitions.*

DEFINITION 3. *an* elementary path *is a path without symbolic repetitions or containing symbolic repetitions in which symbols '+' and '*' are replaced by integers.*

For instance: $(4\ 3), \emptyset, (4\ 5\ 7)^5(4\ 3)^2(4\ 5\ 7)^2 4\ 3$ are elementary paths.

DEFINITION 4. *Given a vertex v, a* raw path *contained in v is a symbolic path beginning from the starting vertex and ending in v. Symbolic repetitions contained in such paths must be effective.*

Paths can be decomposed in two parts starters and sequels.

DEFINITION 5. *Given a raw path ending in v, its* starter *is the largest prefix of this path ending in v and without cycle including v.*

DEFINITION 6. *Given a raw path ending in v, its* sequel *is a cycle ending in v. It corresponds to the remaining part of the raw path when the starter is removed.*

For instance, if we consider the raw path: $(1(7\ 4\ 5)^+3(9\ 8)^+)$, $v = 8$, the starter is $(1(7\ 4\ 5)^+3\ 9\ 8$ and the sequel is $(9\ 8)^+$.

2. DISTRIBUTED SOLVER BASED ON MULTI-CASTES ANT SYSTEM

Solving this problem consists in finding *all* the paths verifying relations (1), (2) and (3). Let us call this set of paths S^*. As this problem is strongly NP-Hard, in reasonnable time it is only possible to determine a subset of S^*.

For that purpose we need a method allowing a careful analysis of the whole graph by visiting many times vertices and by building many paths. The natural collective behavior which exists in social insects population, seems to be particularly well suited for this task [7]. In such system, the consideration of several functions within ants population is biologically expressed by the existence of castes in ants species. Many natural social systems are composed with entities that have simple capabilities even if the collective behavior is complex. So, complex collective behavior emerging from the behavior and interactions of many agents [11] seems to be particulary well suited to resolve the graph exploration under constraints.

Eperimental studies show [8] that natural ants continously forage their territories to search food. These explorations can be described by the search of a path in graph. So ant behavior based algorithms have been succesfully applied in the contex of optimization problems like the Traveling Salesman Problem [5], graph coloring [4] or routing in communication networks [3], for example.

Our approach, is not to find the best path. So it is not typically an optimization problem but we have to enumerate paths that respect contraints. Like in the ant algorithms we use multiple interactive swarms to solve this graph problem.

For that purpose, we propose and describe in the sequel a distributed multi-castes ant system called DIMANTS. The whole system is made of an environment and several populations as illustrated on Figure 2.

Due to the distributed nature of the considered target architecture, we consider the duplication of the environment. The size of considered graphs is not large enough to justify their splitting.

2.1 Environment

The underlying structure of the environment is the SBH-graph. Each vertex contains the symbolic paths that were left by ants in this node. These raw paths are left by some ants and are decomposed by some other ants into a formulation starter and sequel as illustrated in the table 2.

It presents some raw paths among those that were left inside vertices 1, 2 and 3 (Figure 3) and their transformation in a splitted form (starters and sequels of the table were obtained after a simplification process described in Section 2.2.2).

This formulation presents a big interest in the knowledge of what has been explored until this point. Indeed, displaying starters and sequels in a matrix allows the expression of

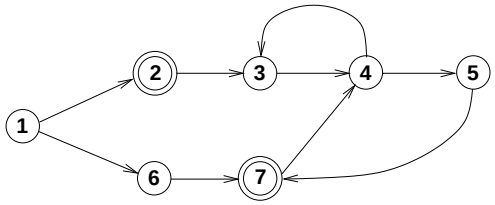

Figure 3: A SBH-graph $G = (V, A)$. $V_a = \{1, 3, 4, 5, 6\}$ and $V_b = \{2, 7\}$.

unexplored potential paths from what is already known. For instance for vertex 3, cells marked by a star correspond to complete paths that have already been explored, while cells containing '?' indicates that no ant put this path in the vertex yet.

	$(4\ 3)^*$	$(4\ 5\ 7\ (4\ 3)^+)^*$
1 2 3	$*$	?
1 6 7 $(4\ 5\ 7)^*$ 4 3	$*$	$*$

The vertices are active elements of the structure. They have some attributes in relation with paths, and they are able to launch some mechanisms by means of signals. The attributes are: raw paths, sets of starters and sequels, the associated potential paths matrix, and a set of thresholds:

- the upper limit of the acceptable number of raw paths contained in the vertex: $\bar{R}$,
- the upper limit of the acceptable number of unexplored paths contained in the potential paths matrix: $\bar{U}$,
- the upper limit of the acceptable quantity of information not communicated to other environments yet $\bar{C}$.

When the threshold $\bar{R}$ is crossed, the vertex sends a signal for indicating that it has to be cleaned, and this signal may affect the current or the next visiting ant. When the threshold $\bar{U}$ is crossed, the vertex sends a signal for indicating that new ants are needed for exploring these paths. This signal may affect the current ant. If the threshold $\bar{C}$ is crossed, the vertex initiates a communication with corresponding vertices belonging to other copies of the environment.

2.2 Population

A population is characterized by a set of attributes that will be detailed in the sequel. The population is composed of individuals belonging to several castes. We have considered three different castes for our problem, *explorers*, *cleaners* and *reproductives*.

2.2.1 Explorer

One ant belonging to this caste is able to walk through the graph and to build a path. As soon as it visits a vertex, this latter is inserted in the current path. The way the vertex is inserted in the path depends on the context. If the vertex is not in the path yet, it is just concatenated to the current path. If the vertex is already in the path and belongs to

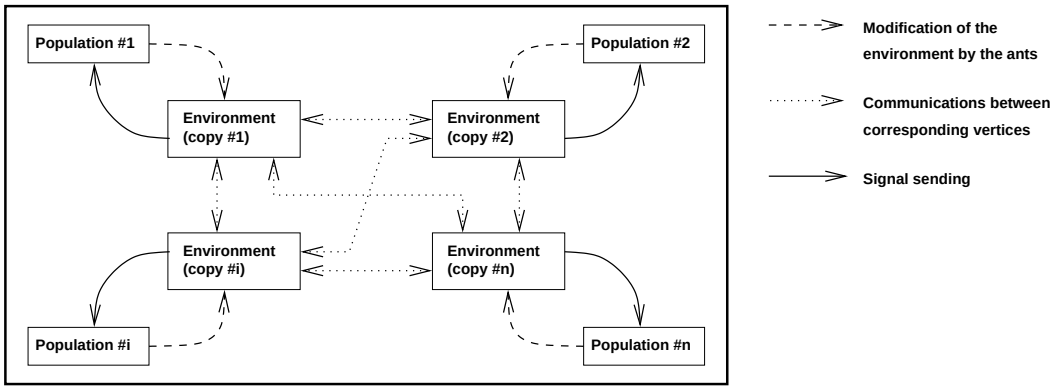

Figure 2: A synthetic view of the system.

Vertices	Raw paths	Starters	Sequels
1	1	1	$\emptyset$
2	1 2	1	2
3	1 2 3 1 2 (3 4)$^+$ 3 1 6 7 4 3 1 6 7 (4 3)$^+$ 1 6 (7 4 5)$^+$7 4 3 1 6 (7 4 5)$^+$ 7 (4 3)$^+$ 1 6 7 (4 3)$^+$ (4 5 7 (4 3)$^+$)$^+$	1 2 3 1 6 7 (4 5 7)* 4 3	(4 3)* (4 5 7 (4 3)$^+$)*

Table 2: Raw paths transformation in splitted form

a loop, it is just added at the end of the path unless the sequence of vertices following the loop is identical to the loop itself in which case this sequence is removed. Example of paths of length 9 built from the graph of Figure 4 by three different explorers are given in the table below.

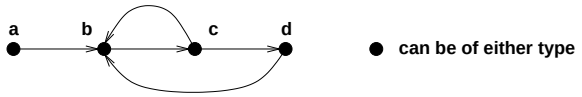

Figure 4: An example of graph with four vertices of whatever type containing two cycles (bc) **and** (bcd)**.**

After the addition of the vertex to the path, a verification is performed. If among relations (1), (2) and (3) only one is not verified, the ant dies. Otherwise, before moving to another vertex the ant leaves a copy of the current path. The next node has to be chosen between the set of successors (denoted by S_{succ}) of the current vertex. Different strategies are possible for performing this choice.

An explorer can become a cleaner in response to a signal coming from the environment.

2.2.2 Cleaner

The role of cleaners consists in transforming raw paths into splitted formulations as described in Section 2.1. From the splitted formulations (Starters, Sequels), the cleaner builds the potential paths matrix.

The objectives of the cleaning process are to remove redundant paths, to compress path formulation and to determine new symbolic paths by combining starters with series of sequels. The method is composed of three steps: decomposition of raw paths into starters and sequels, processing of starters and processing of sequels. The decomposition is performed according to the definitions 5 and 6. Redundant starters are removed during the second step of the cleaning process as well as starters that can be generated by other starters. During the last step, redundant sequels are removed and if $\emptyset$ belongs to the set of sequels, all effective symbolic repetitions are transformed into general ones. If a sequel is made of a concatenated loops, the method replaces the sequel by the loops themselves.

A cleaner is stemmed from an explorer affected by a signal coming from the visited vertex. When the task of the cleaner is completed, and if no signal is coming from the environment, the cleaner becomes again an explorer.

2.2.3 Reproductive

From the vertex in which it is located, a reproductive creates new ants all of the explorer caste. Each reproductive produces ants according to its own set of parameters: the percentage of each strategy, and the set of probabilities for a change of caste in response to a signal sent by the environment.

At the very beginning, one reproductive is launched in the starting vertex. One such reproductive with its own parameters set is created in every copy of the environment. The set of parameters owned by the reproductive is transmitted to the created ants.

htp

vertex	path	vertex	path	vertex	path
a	a	a	a	a	a
b	ab	b	ab	b	ab
c	abc	c	abc	c	abc
b	$a(bc)^+b$	d	$abcd$	b	$a(bc)^+b$
c	$a(bc)^+$	b	$a(bcd)^+b$	c	$a(bc)^+$
d	$a(bc)^+d$	c	$a(bcd)^+bc$	d	$a(bc)^+d$
b	$a(bc)^+db$	b	$a(bcd)^+(bc)^+b$	b	$a(bc)^+db$
c	$a(bc)^+dbc$	c	$a(bcd)^+(bc)^+$	c	$a(bc)^+dbc$
b	$a(bc)^+d(bc)^+b$	d	$a(bcd)^+(bc)^+d$	d	$a(bc)^+(dbc)^+d$

2.3 Interactions Environment ↔ Populations

There exist some interactions between the environment and the individuals. Individuals leave some information in the environment which sends signals to individuals for special tasks, cleaning, request for exploration, request for communications. For that, the environment is able to change the caste of visiting ants by means of signals. This follows adaptive complex systems model where auto-organizations come from feed-back expressed here by the environment.

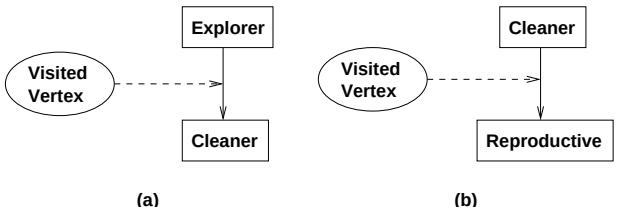

Figure 5: Signals sent by the environments may change the caste of some ants. When the threshold $\bar{R}$ of the visited vertex is crossed, this one sends a signal that may change the caste of the current ant from explorer to cleaner (a). When the threshold $\bar{U}$ is crossed, the current visited vertex sends a signal that may imply the change of the caste of the ant from cleaner to reproductive (b).

3. CONCLUSION AND PERSPECTIVES

Experiments are in development and we are going to study the rate of solutions founded by seconds and the scability in term of processes number. This method presents some interesting and precious qualities:

- after the production of the first element belonging to S^*, the algorithm becomes anytime,
- given that the SBH-graph is duplicated on each processor, a failure of one or more processing element does not prevent the process to be carried out,
- the frequency of exchanges of information can be adapted to the set of characteristics of the communication environment (transmission rates, links capacity, topologies (diameter, connexity...)).

4. REFERENCES

[1] W. Bains and G. C. Smith. A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology*, 135:303–307, 1988.

[2] J. Błażewicz, P. Formanowicz, M. Kasprzak, W. T. Markiewicz, and J. Węglarz. DNA sequencing with positive and negative errors. *Journal of Computational Biology*, 6:113–123, 1999.

[3] G. D. Caro and M. Dorigo. Antnet: A mobile agents approach to adaptive routing. Technical report, IRIDIA, Université libre de Bruxelles, Belgium, 1997.

[4] D. Costa and A. Hertz. Ant can colour graphs. *JORS*, (48):295–305, 1997.

[5] M. Dorigo and L. Gambardella. Ant colony system: A cooperative learning approach to the traveling saleman problem. *IEEE Transactions on Evolutionary Computation*, 1(1):53–66, 1997.

[6] R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov. Sequencing of megabase plus DNA by hybridization: theory and the method. *Genomics*, 4:114–128, 1989.

[7] M. D. E. Bonabeau and G. Theraulaz. *Swarm Intelligence - From natural to Artificial Systems*. Oxford University Press, 1999.

[8] D. Gordon. The expandable network of ant exploration. *Animal Behaviour*, 50:995–1007, 1995.

[9] A. Guénoche. Can we recover a sequence, just knowing all its subsequences of given length? *Computer Applications in the Biosciences (CABIOS)*, 8:569–574, 1992.

[10] F. Guinand, L. Mouchard, and A. Rabia. Impact of repetitions on sbh models. Technical Report P-01-02, LIH, Le Havre - France, 2002.

[11] C. Langton, editor. *Artificial Life*. Addison Wesley, 1987.

[12] Lysov, P. Yu, V. L. Florentiev, A. A. Khorlin, K. R. Khrapko, V. V. Shik, and A. D. Mirzabekov. Determination of the nucleotide sequence of DNA using hybridization with oligonucleotides. *Doklady Akademii Nauk SSSR*, 303:1508–1511, 1988.

[13] P. Pevzner. l-tuple DNA sequencing: computer analysis. *Journal of Biomolecular Structure and Dynamics*, 7:63–73, 1989.