Molecular Computation Of Solutions To Combinatorial Problems

Leonard M. Adleman*

Professor Leonard M. Adleman

Department of Computer Science

and

Institute for Molecular Medicine and Technology

University of Southern California

941 west 37th place

Los Angeles, CA 90089-0781

U.S.A.

Research Supported By National Science Foundation (Grant #CCR-9214671)

The tools of molecular biology are used to solve an instance of the directed Hamiltonian path problem. A small graph is encoded in molecules of DNA and the 'operations' of the computation are performed with standard protocols and enzymes. This experiment demonstrates the feasibility of carrying out computations at the molecular level.

In 1959 Richard Feynman gave a visionary talk describing the possibility of building computers which were "*sub*-microscopic" (1). Despite remarkable progress in computer miniaturization this goal has yet to be achieved. In this report the possibility of computing directly with molecules is explored.

A directed graph G with designated vertices v_{in} and v_{out}, is said to have a Hamiltonian path (2) if and only if there exists a sequence of compatible 'one way' edges $e_1, e_2, ..., e_z$ (that is, a 'path') which begins at v_{in}, ends at v_{out} and enters every other vertex exactly once. Figure 1 shows a graph which for $v_{in} = 0$ and $v_{out} = 6$ has a Hamiltonian path, given by the edges $0 \rightarrow 1, 1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4, 4 \rightarrow 5, 5 \rightarrow 6$. If the edge $2 \rightarrow 3$ were removed from the graph then the resulting graph with the same designated vertices would not have a Hamiltonian path. Similarly if the designated vertices were changed to $v_{in} = 3$, $v_{out} = 5$ there would be no Hamiltonian path (since for example there are no edges entering vertex 0).

There are well known algorithms for deciding whether an arbitrary directed graph with designated vertices has a Hamiltonian path or not. However, all known algorithms for this problem have exponential worst-case complexity and hence there are instances of modest size for which these algorithms require an impractical amount of computer time to render a decision. Since the directed Hamiltonian path problem has been proven to be NP-complete, it seem likely that no efficient (that is, polynomial time) algorithm exists for solving it (2,3).

The following (non-deterministic) algorithm solves the directed Hamiltonian path problem:

- Step 1: Generate random paths through the graph.
- Step 2: Keep only those paths which begin with v_{in} and end with v_{out}.
- Step 3: If the graph has n vertices, then keep only those paths which enter exactly n vertices.
- Step 4: Keep only those paths which enter all of the vertices of the graph at least once.
- Step 5: If any paths remain, say "YES", otherwise say "No".

The graph shown in Figure 1 with designated vertices $v_{in} = 0$ and $v_{out} = 6$ was solved using the algorithm above implemented at the molecular level. Note that the labeling of the vertices in such a way that the (unique) Hamiltonian path enters the vertices in sequential order, is only for convenience in this exposition and provides no advantage in the computation. The graph is small enough that the Hamiltonian path can be found by visual inspection; however, it is large enough to demonstrate the feasibility of this approach. It seems clear that the methods described here could be scaled-up to accommodate much larger graphs.

To implement Step 1 of the algorithm, each vertex i in the graph was associated with a random 20-mer sequence of DNA denoted O_i. For each edge $i \rightarrow j$ in the graph an oligonucleotide $O_{i\rightarrow j}$ was created which was the 3' 10-mer of O_i (unless $i = 0$ in which case it was all of O_i) followed by the 5' 10-mer of O_j (unless $j = 6$ in which case it was all of O_j). Notice that this construction preserves edge orientation. For example, $O_{2\rightarrow 3}$ will not be the same as $O_{3\rightarrow 2}$. The 20-mer sequence Watson-Crick complementary to O_i was denoted $\overline{O}_i$. Figure 2 contains examples.

For each vertex i in the graph (except $i = 0$ and $i = 6$) 50 pmol of $\overline{O}_i$ and for each edge $i \rightarrow j$ in the graph 50 pmol of $O_{i \rightarrow j}$ were mixed together in a single ligation reaction (4). The $\overline{O}_i$ served as splints to bring oligonucleotides associated with compatible edges together for ligation (see Figure 2). Hence the ligation reaction resulted in the formation of DNA molecules encoding random paths through the graph.

The scale of this ligation reaction far exceeded what was necessary for the graph under consideration. For each edge in the graph, approximately 3×10^{13} copies of the associated oligonucleotides were added to the ligation reaction. Hence it is likely that vast numbers of DNA molecules encoding the Hamiltonian path were created. In theory the creation of a single such molecule would be sufficient. Hence, for this graph, sub-attomol quantities of oligonucleotides would probably have been sufficient. Alternatively, a much larger graph could have been processed with the pmol quantities employed here.

To implement Step 2 of the algorithm, the product of Step 1 was amplified by polymerase chain reaction (PCR) using primers O_0 and $\overline{O}_6$ (5). Thus only those molecules encoding paths which begin with vertex 0 and end with vertex 6 were amplified.

To implement Step 3 of the algorithm, the product of Step 2 was run on an agarose gel and the 140bp band (corresponding to dsDNA encoding paths entering exactly seven vertices) was excised and soaked in ddH_2O to extract DNA (6). This product was PCR amplified and gel purified several times to enhance purity.

To implement Step 4 of the algorithm, the product of Step 3 was affinity purified using a biotin-avidin magnetic beads system. This was accomplished by first generating single stranded DNA from the dsDNA product of Step 3 and then incubating the ssDNA with $\overline{O}_1$ conjugated to magnetic beads (7). Only those ssDNA molecules which contained the sequence O_1 (and hence encoded paths which entered vertex 1 at least once) annealed to the bound $\overline{O}_1$ and were retained. This process was repeated successively with $\overline{O}_2$, $\overline{O}_3$, $\overline{O}_4$ and $\overline{O}_5$.

To implement Step 5, the product of Step 4 is PCR amplified and run on a gel.

Figure 3 shows the results of these procedures. In panel A, lane 1 is the result of the ligation reaction in Step 1. The smear with striations is consistent with the construction of molecules encoding random paths through the graph (8). Panel A, lanes 2-5 show the results of the PCR reaction in Step 2. The dominant bands correspond to the amplification of molecules encoding paths which begin at vertex 0 and end at vertex 6.

Panel B shows the results of a 'graduated PCR' performed on the ssDNA molecules generated from the band excised in Step 3. Graduated PCR is a method for 'printing' results. Graduated PCR is performed by running 6 different PCR reactions using as right primer O_0 and left primer $\overline{O_i}$ in the i^{th} tube. For example, on the molecules encoding the Hamiltonian path 0→1,1→2,2→3,3→4,4→5,5→6, graduated PCR will produce bands of 40bp,60bp,80bp,100bp,120bp,140bp in successive lanes. On the molecules encoding the path 0→1,1→3,3→4,4→5,5→6, graduated PCR will produce bands of 40bp,x,60bp,80bp,100bp,120bp in successive lanes where the x denotes the absence of a band in lane 2 (corresponding to the omission of vertex

2 along this path). On molecules encoding the path 0→3,3→2,2→3,3→4,4→5,5→6, graduated PCR will produce bands of x,60bp,80bp/40bp,100bp,120bp,140bp in successive lanes, where 80bp/40bp denotes that both a 40bp and an 80bp band will be produced in lane 3 (corresponding to the double passage of vertex 3 along this path). The most prominent bands in Panel B appear to be those which would arise from superimposing the bands predicted for the three paths described above. The bands corresponding to path 0→1,1→3,3→4,4→5,5→6, were not expected and suggest that the band excised in Step 3 contained contamination from 120bp molecules. However, such low weight contamination is not a problem since it does not persist through Step 4.

Panel C shows the results of graduated PCR applied to the molecules in the final product of Step 4. The bands demonstrate that the these molecules encode the Hamiltonian path 0→1,1→2,2→3,3→4,4→5,5→6 (9).

The computation above required approximately 7 days of lab work. Step 4 (magnetic bead separation) was the most labor intensive, requiring a full day at the bench. In general, using the algorithm above, the number of procedures required should grow linearly with the number of vertices in the graph. The labor required for large graphs might be reduced by using alternative procedures, automation or less labor intensive molecular algorithms.

The number of different oligonucleotides required should grow linearly with the number of edges. The quantity of each oligonucleotide needed is a rather subtle graph theoretic question (8). Roughly, the quantity used should be just sufficient to insure that during the ligation step (Step 1) a molecule encoding a Hamiltonian path will be formed with high probability if such a path exists in the graph. This quantity should grow exponentially with the number of vertices in the graph. The molecular algorithm used here was rather naive and as with classical computation, finding improved algorithms will extend the applicability of the method.

As the computation is scaled up, the possibility of errors will need to be looked at carefully. During Step 1, the occasional ligation of incompatible edge oligonucleotides may result in the formation of molecules encoding 'pseudo paths' which do not actually occur in the graph. While such molecules may be amplified during Step 2 and persist through Step 3, they seem unlikely to survive the separation in Step 4. Nonetheless, at the completion of a computation, it would be prudent to confirm that a putative Hamiltonian path actually occurs in the graph. During the separation step, molecules encoding Hamiltonian paths may fail to bind adequately and be lost, while molecules encoding non-Hamiltonian paths may bind non-specifically and be retained. The latter problem might be mitigated by more stringent or repeated separation procedures. The former problem might be dealt with by periodically applying PCR with primers designed to amplify Hamiltonian paths (in the example above primers O_0 and $\overline{O_6}$). The balanced use of these techniques may be adequate to control such errors.

The choice of random 20-mer oligonucleotides for encoding the graph was based on the following rationale. First, since 4^{20} 20-mer oligonucleotides exist, choosing randomly made it unlikely that oligonucleotides associated with different vertices would share long common subsequences which might result in 'unintended' binding during the ligation step (Step 1). Second, it was guessed that with high probability potentially deleterious (and presumably rare) features such as severe hairpin loops would not be likely to arise. Finally, choosing 20-mers assured that binding between 'splint' and 'edge' oligonucleotides would involve ten nucleotide pairs and would consequently be stable at room temperature. This approach was successful for the small graph considered above; however, how to best proceed for larger graphs may require additional research.

What is the power of this method of computation? It is premature to give definitive answers; however some remarks seem in order.

A typical desk top computer can execute approximately 10^6 operations per second. The fastest super computers currently available can execute approximately 10^{12} operations per second. If the ligation (concatenation) of two DNA molecules is considered as a single operation and if it is assumed that about half of the approximately 4×10^{14} 'edge' oligonucleotides in Step 1 were ligated, then during Step 1 approximately 10^{14} operations were executed. Clearly, this step could be scaled-up considerably and 10^{20} or more operations seems entirely plausible (for example by using μmol rather than pmol quantities). At this scale, the number of operations per second during the ligation step would exceed that of current super computers by more than a thousand fold. Further, hydrolysis of a single molecule of ATP to AMP plus pyrophosphate provides the energy ($\Delta G = -8$ kcal/mol) for one ligation operation (10,11); hence in principle 1 joule is sufficient for approximately 2×10^{19} such operations. This is remarkable energy efficiency, considering that the second law of thermodynamics dictates a theoretical maximum of 34×10^{19} (irreversible) operations per joule (at 300° K) (12,13). Existing super computers are far less energy efficient; executing at most 10^9 operations per joule. The energy consumed during other parts of the molecular computation such as oligonucleotide synthesis and PCR should also be small in comparison to that consumed by super computers. Finally, storing information in molecules of DNA allows for an information density of approximately 1 bit per cubic nm, a dramatic improvement over existing storage media such as video tape which store information at a density of approximately 1 bit per 10^{12} cubic nanometers. Thus the potential of molecular computation is impressive. What is not clear is whether such massive numbers of inexpensive operations can be productively used to solve real computational problems. One major advantage of electronic computers is the variety of operations they provide and the flexibility with which these operations can be applied. While two 100 digit integers can be multiplied quite efficiently on an electronic computer; it would be a daunting task to do such a calculation on a molecular computer using currently available protocols and enzymes (14).

Nonetheless, for certain intrinsically complex problems, such as the directed Hamiltonian path problem where existing electronic computers are very inefficient and where massively parallel searches can be organized to take advantage of the operations that molecular biology currently provides, it is conceivable that molecular computation might compete with electronic computation in the near term. It is a research problem of considerable interest to elucidate the kinds of algorithms which are possible using molecular methods and the kinds of problems which these algorithms can efficiently solve (12,15,16).

For the long term one can only speculate about the prospects for molecular computation. It seems likely that a single molecule of DNA can be used to encode the 'instantaneous description' of a Turing machine (17) and that currently available protocols and enzymes could (at least under idealized conditions) be used to induce successive sequence modifications which would correspond to the execution of the machine. In the future, research in molecular biology may provide improved techniques for manipulating macromolecules. Research in chemistry may allow for the development of synthetic 'designer' enzymes. One can imagine the eventual emergence of a general purpose computer consisting of nothing more than a single macromolecule conjugated to a ribosome-like collection of enzymes which act upon it.

References and Notes

References

[1] R.P. Feynman, in *Minaturization*, D.H. Gilbert, Ed. (Reinhold Publishing Corporation, New York, 1961), pp. 282-296.

[2] M.R. Garey and D.S. Johnson, *Computers and Intractability* (W.H. Freeman and Co., San Francisco, CA, 1979).

[3] R.M. Karp, in *Complexity of Computer Computations*, R.E. Miller and J.W. Thatcher, Eds. (Plenum Press, New York, NY, 1972), pp. 85-103.

[4] 50 pmol of each oligonucleotide with 5'-terminal phosphate residue, 5 units T4 DNA ligase (Boehringer-Mannheim, Germany), ligase buffer and ddH_2O to a total volume of 100μl were incubated for 4 hours at room temperature.

[5] All PCR amplifications were performed on a Perkins-Elmer (Norwalk, CT) 9600 thermal cycler. For amplification in Step 2: 50 pmol of each primer, 5 units Taq DNA polymerase (Gibco-BRL,

Grand Island, NY) in PCR buffer to a total volume of 50μl were processed for 35 cycles of 94°C, 15 seconds, 30°C, 60 seconds. For graduated PCR: 50 pmol of each primer, 2.5 units Taq DNA polymerase in PCR buffer to a total volume of 50μl were processed for 25 cycles of 94°C, 15 seconds, 40°C, 60 seconds.

[6] All gels were 3% or 5% agarose (NuSieve, FMC BioProducts, Rockland, ME) in TBE buffer with ethidium bromide staining (14).

[7] Oligonucleotides were 5' biotinylated using LC Biotin-ON Phosphoramidite (Clontech). To obtain ssDNA, the product from Step 3 was PCR amplified using primers O_0 and biotinylated $\overline{O}_6$. The amplified product was annealed to Streptavidin Paramagnetic Particles (Promega, Madison, WI) by incubating in 100 μl 0.5x SSC for 45 minutes at room temperature with constant shaking. Particles were washed 3 times in 200μl of 0.5x SSC and then heated to 80°C in 100 μl ddH_2O for 5 minutes to denature the bound dsDNA. The aqueous phase with ssDNA was retained. For affinity purification, 1 nmol biotinylated $\overline{O}_1$ was annealed to par-

ticles as above and washed 3 times in 400μl of 0.5x SSC. ssDNA was then incubated with these particles in 150μl 0.5x SSC for 45 minutes at room temperature with constant shaking. Particles were washed 4 times in 400μl of 0.5x SSC to remove unbound ssDNA and then heated to 80°C in 100 μl ddH_2O for 5 minutes to release ssDNA bound to $\overline{O}_1$. The aqueous phase with ssDNA was retained. This process was then repeated for $\overline{O}_2$, $\overline{O}_3$, $\overline{O}_4$ and $\overline{O}_5$.

[8] From a graph theoretic point of view, the use of equal quantities of each oligonucleotide in the ligation reaction is not optimal and leads to the formation of excess numbers of molecules encoding paths which do not start at vertex 0 or do not end at vertex 6. A better way to proceed is to first calculate a flow on the graph and use the results to determine the quantity of each oligonucleotide that is necessary.

[9] On an n vertex graph G with designated vertices v_{in} and v_{out} there may be multiple Hamiltonian paths. If it is desirable to have an explicit description of some Hamiltonian path, that can be accomplished by extending the algorithm as follows. At the end

of step 4 one has a solution (in the chemistry sense) containing molecules encoding all Hamiltonian paths for $< G, v_{in}, v_{out} >$. The graduated PCR performed at the end of step 4 will produce the superimposition of the bands corresponding to all of these Hamiltonian paths in the $n-1$ successive lanes. For some lane i, a band of least weight (40bp) will appear. This indicates that some Hamiltonian path begins with v_{in} and proceeds directly to vertex i. By PCR amplifying the solution with primers O_i and $\overline{O}_n$, running a gel and excising the $20*(n-1)$bp band, only those molecules encoding such Hamiltonian paths will be retained. One now has a solution containing molecules encoding all Hamiltonian paths for $< G'$,i$, v_{out} >$ where G' is the graph where vertex v_{in} has been removed. One now iterates.

[10] J.D. Watson, N.H. Hopkins, J.W. Roberts, J.A. Steitz and A.M. Weiner, *Molecular Biology of the Gene* (The Benjamin/Cummings Publishing Co., Menlo Park, CA, ed. 3, 1987).

[11] M.J. Engler and C.C. Richardson, in *The Enzyme* P.D. Boyer, Ed. (Academic Press Inc., New York, NY , ed. 3, 1982) vol. XVB pp.

3.

[12] T.D. Schneider, *J. theor. Biol.*, **148**, 125 (1991).

[13] R.C. Merkle, *Nanotechnology*, **4**, 21-40 (1993).

[14] J. Sambrook, E.F. Fritsch and T. Maniatis, *Molecular Cloning* (Cold Spring Harbor Laboratory Press, Cold Springs Harbor, New York, ed. 2, 1989).

[15] B.C. Crandall and J. Lewis, Eds. *Nanotechnology* (MIT Press, Cambridge, MA, 1992).

[16] D. Bradley, *Science*, **259**, 890 (1993).

[17] H. Rogers Jr. *Theory of Recursive Functions and Effective Computability* (McGraw-Hill Book Company, New York, 1967).

[18] The author would like to express his gratitude to Dr. Chelypov for teaching him molecular biology. The author would also like to thank Salahuddin for making the resources of the Institute for Molecular Medicine And Technology available. The author thanks Dr. Deonier for helpful discussions. Research supported in part by

the National Science Foundation (Grant #CCR-9214671) and a Zumberg Research Initiation Fund grant from the University of Southern California.

FIGURE LEGENDS

Figure1: Directed graph. When $v_{in} = 0$ and $v_{out} = 6$ a unique Hamiltonian path exists: 0→1,1→2,2→3,3→4,4→5,5→6.

Figure 2: Encoding a graph in DNA.. For each vertex i in the graph, a random 20-mer oligonucleotides O_i is generated (shown: O_2, O_3, O_4 for vertices 2,3,4 respectively). For edge $i \rightarrow j$ in the graph an oligonucleotide $O_{i \rightarrow j}$ is derived from the 3' 10-mer of O_i and the 5' 10-mer of O_j (shown: $O_{2 \rightarrow 3}$ for edge 2→3, $O_{3 \rightarrow 4}$ for edge 3→4). For each vertex i in the graph $\overline{O}_i$, is the Watson-Crick complement of O_i (shown: $\overline{O}_3$, the complement of O_3). Notice $\overline{O}_3$ serves as a splint to bind $O_{2 \rightarrow 3}$ and $O_{3 \rightarrow 4}$ in preparation for ligation. All oligonucleotides written 5' to 3' except $\overline{O}_3$.

Figure 3: Agarose gel electrophoresis of various products of the experiment. Panel A: product of ligation reaction (lane 1), PCR amplification of product of ligation reaction (lanes 2-5), molecular weight marker (lane 6). Panel B: graduated PCR of product from Step 3 (lanes 1-6), molecular weight marker

(lane 7). Panel C: graduated PCR of the final product of the experiment revealing Hamiltonian path (lanes 1-6), molecular weight marker (lane 7).