Protein output for DNA computing

Henkel, C.V.; Bladergroen, R.S.; Balog, C.I.A.; Deelder, A.M.; Head, T.; Rozenberg, G.; Spaink, H.P.

Nat Comput (2005) **4**:1–10
DOI: 10.1007/s11047-004-5199-x

Protein output for DNA computing

CHRISTIAAN V. HENKEL [1,2,3,*], RENO S. BLADERGROEN [1,2], CRINA I. A. BALOG [4], ANDRÉ M. DEELDER [4], TOM HEAD [5], GRZEGORZ ROZENBERG [1,3] and HERMAN P. SPAINK [1,2]
[1]*Leiden Center for Natural Computing, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands;* [2]*Institute of Biology, Leiden University, Wassenaarseweg 64, 2333 AL Leiden, The Netherlands;* [3]*Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands;* [4]*Department of Parasitology, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands;* [5]*Department of Mathematical Sciences, Binghamton University, Binghamton NY 13902-6000, USA; (*Author for correspondence, e-mail: henkel@rulbim.leidenuniv.nl)*

Abstract. In recent years, several strategies for DNA based molecular computing have been investigated. An important area of research is the detection and analysis of output molecules. We demonstrate how DNA computing can be extended with *in vivo* translation of the output. In the resulting proteins, the information per kilogram is about 15-fold higher than in the original DNA output. The proteins are therefore of correspondingly smaller mass, which facilitates their subsequent detection using highly sensitive mass spectrometry methods. We have tested this approach on an instance of the Minimal Dominating Set problem. The DNA used in the computation was constructed as an open reading frame in a plasmid, under the control of a strong inducible promoter. Sequential application of restriction endonucleases yielded a library of potential solutions to the problem instance. The mixture of plasmids was then used for expression of a protein representation. Using MALDI-TOF mass spectrometry, a protein corresponding to the correct solution could be detected. The results indicate the feasibility of the extension of DNA computing to include protein technology. Our strategy opens up new possibilities for both scaling of DNA computations and implementations that employ output of functional molecules or phenotypes.

Key words: DNA computing, plasmid computing, proteomics

1. Introduction

Molecular computers based on DNA were originally proposed as an alternative or supplement to electronic computers (Adleman, 1994; Jonoska and Seeman, 2002; Hagiya and Ohuchi, 2003). Features such as

the inherent logic of DNA hybridization and massive parallelism are very interesting from the computational point of view. DNA computers have the potential to extend the range of solvability for computationally hard problems (Paun et al., 1998; Jonoska and Seeman, 2002; Hagiya and Ohuchi, 2003). A number of experiments solving small-scale instances of well known computational problems have since been described (Ouyang et al.,1997; Liu et al., 2000; Sakamoto et al., 2000; Mao et al., 2000; Benenson et al., 2001; Braich et al., 2002). Today, DNA is also under investigation as a component of special purpose computers (Normile, 2002) and as a storage medium for non-natural information (Cox, 2001).

An important problem in biomolecular computing is the generation and analysis of output molecules. Here, we have exploited the natural capacity of DNA to direct the synthesis of proteins, which can be used as output molecules. An advantage of the use of proteins for molecular computations is that much higher information densities are possible than using nucleic acids: using translation, the information content of a DNA triplet (approximately 2000 Da) is expressed in one 57–186 Da amino acid. The small proteins can then be accurately analyzed using modern proteomics technology, where the original DNA molecules would be too bulky to examine by mass spectrometry.

A computation was conducted on a DNA sequence constituting an open reading frame (ORF), which was placed under the control of a strong promoter (Figure 1). This enables the *in vivo* transcription and translation of the computational construct into a protein. Mass spectrometry then allows sensitive determination of both size and composition of the expressed library in parallel.

We have tested the principle of protein output on an instance of the Minimum Dominating Set (MDS) problem, using plasmid DNA as computing hardware (Head et al., 2000). In this approach, restriction endonucleases are used to specifically remove 'stations' from a computing plasmid (Figure 1). The absence or presence of these stations in the plasmid, which is basically a computer memory, corresponds to a bit set to either 0 or 1.

The computation starts with an aqueous solution of a single species of plasmid. Because of the fluid medium and the high number of plasmids, it can be assumed that splitting the mixture in several distinct volumes yields as many identical copies of the memory. These memories are then modified by application of certain restriction endonucleases and religation (removal of stations). This enzymatic 'software' acts on millions of plasmids in parallel. After memory writing, the subsets are mixed again. Iteration of this procedure results in a library containing

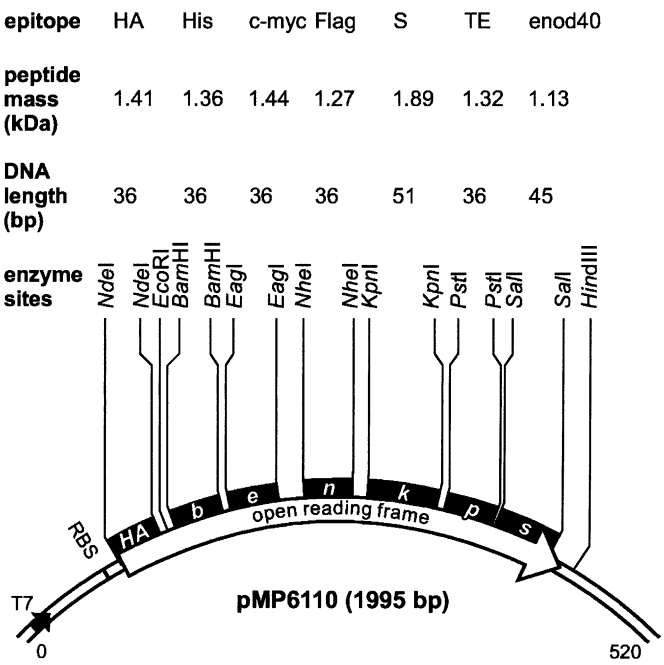

Figure 1. The computing region of computational plasmid pMP6110. Information is present at different levels: a 'station' in the plasmid has a certain length on DNA level, and the associated peptide has a certain weight. Sequences have been chosen such that DNA length corresponds to protein weight. Additionally, the DNA sequences encode known protein epitopes. For these, antibodies are commercially available, enabling an alternative detection system. One of these peptide tags (HA) is used for purification of the expressed protein library and is not used in the computation. Stations are named after the enzymes that can excise them.

an exponential number of plasmids. DNA representing a solution to a given computational problem can be identified by selection on certain characteristics, for example length. Plasmid computing was successfully used to solve a 6-bit instance of the NP-complete Maximal Clique problem (Head et al., 2000) and can be adapted to a broad range of algorithmic problems (Head et al., 2002).

2. Materials and methods

2.1. *Computational plasmid*

Plasmid pMP6110 is a length-minimized derivative of pOK12 (Vieira and Messing, 1991; Head et al., 2002), containing an *E. coli* origin of

replication and a kanamycin resistance marker. The computing ORF is under the control of the strong inducible T7 promoter and a consensus ribosome binding sequence (RBS). Stations in the ORF (Figure 1) encode the following protein epitopes: HA, 6x His-tag and c-myc (Roche Diagnostics); Flag (Sigma-Aldrich); S-tag (Novagen); TE (thrombin and enterokinase cleavage sites); *enod40* (Staehelin et al., 2001). Synthetic oligonucleotides used in plasmid construction were purchased from Isogen Bioscience (Maarssen, NL).

2.2. *Library generation*

An initial plasmid quantity of 40 ng was sequentially digested and religated as shown in Figure 2b. Enzymes were purchased from New England Biolabs and handled according to the manufacturers recommendations. After all enzymatic reactions, the reaction mixture was purified using QIAquick PCR cleanup kits (Qiagen). Ligations were carried out overnight at 16 °C in a 400 μl reaction volume. This combination of fragment removal and large volume minimizes the likelihood of religation of the excised station. After ligation, the plasmid mixture was transformed into *E. coli* XL1-blue cells for amplification and isolated again using QIAprep plasmid miniprep kits (Qiagen). The resulting library was analysed using polyacrylamide gel electrophoresis and bands were visualized using SYBR Green (Molecular Probes, Leiden, NL).

2.3. *Protein purification*

Protein was purified from *E. coli* BL21(DE3) (Invitrogen) induced with isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were lysed in 8 M urea, extract was dialysed against 10 mM Tris pH 8, 1 mM EDTA and tagged proteins were purified using an anti-HA affinity column (Roche Diagnostics). Purified protein was concentrated using Microcon YM-3 concentrators (Millipore). Gel electrophoresis and staining were performed as described (Schagger and Von Jagow, 1987; Sambrook and Russell, 2001).

2.4. *Mass spectrometry*

Protein samples were desalted using Centri Spin 10 gel filtration columns (emp Biotech, Berlin). Spectra were recorded on a Bruker

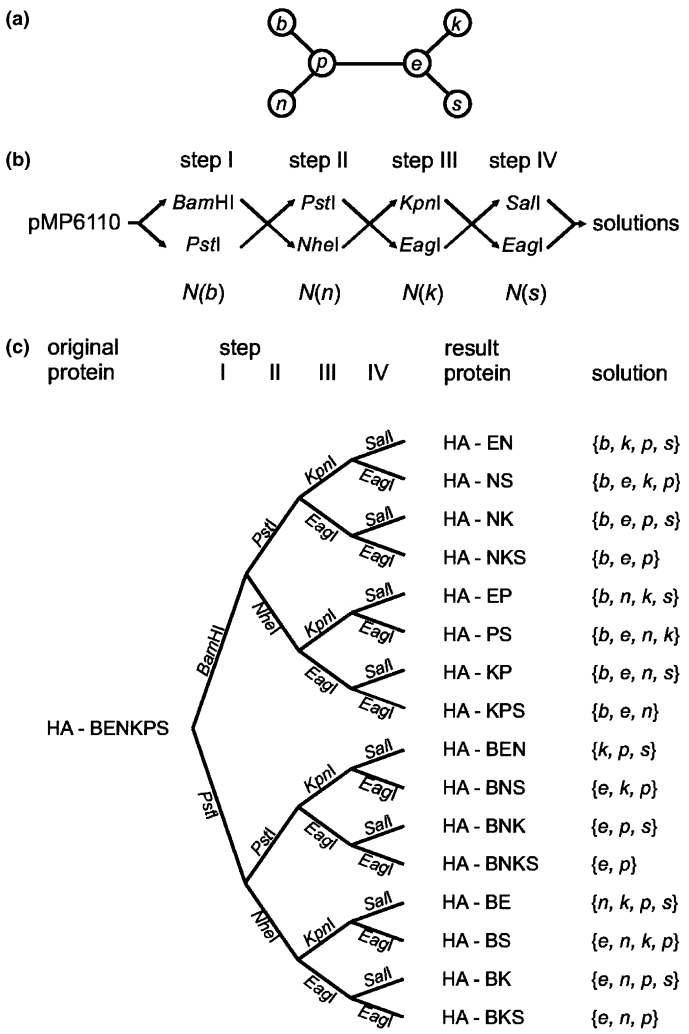

Figure 2. Problem instance and solution strategy. (a) Schematic representation of the graph used. Vertices are named after the available 'stations' on plasmid pMP6110 (Figure 1). The graph can be defined as an undirected graph $G = (V, E)$, with V the set of vertices and E the set of edges. Alternatively, the graph can be defined by a set of neighbourhoods: a neighbourhood $N(v)$ for vertex v is defined as the set $N(v) = \{u \text{ in } V$: either $u = v$ or $[u, v]$ is in $E\}$. The graph shown here is then given by $N(b) = \{b, p\}$, $N(e) = \{e, k, p, s\}$, $N(n) = \{n, p\}$, $N(k) = \{e, k\}$, $N(p) = \{b, e, n, p\}$, $N(s) = \{e, s\}$. A subset of V is a dominating set precisely in the case that it contains at least one vertex from every neighbourhood. (b) Generation of all possible solutions by digestion and ligation. Only four steps are necessary, since neighbourhoods $N(e)$ and $N(p)$ are redundant. For example, $N(e)$ contains $N(k)$ and $N(s)$, and is therefore already accommodated in step III or step IV. (c) The complete library generated by the procedure shown in Figure 2b.

Reflex III mass spectrometer in linear mode, using 2,5-dihydroxybenzoic acid supplemented with fucose as a matrix. The identity of the original pMP6110 protein product was confirmed by analysis of trypsin and chymotrypsin digests on a Bruker Ultraflex (data not shown).

3. Results

3.1. *Minimal dominating set*

Given a graph with vertices (nodes) and edges (connections), the MDS problem asks for the smallest possible vertex set from which all other vertices can be reached by an edge. The graph used is shown in Figure 2a.

The MDS problem is a representative of the large and important class of NP-complete problems and is as such equivalent to all other problems in this class (Garey and Johnson, 1979). Other instances of NP-complete problems that have been used to test DNA computing approaches include Directed Hamiltonian Path (Adleman, 1994), Maximal Clique (Ouyang et al., 1997) and Satisfiability (Faulhammer et al., 2000; Liu et al., 2000; Sakamoto et al., 2000; Braich et al., 2002) problems. In particular, the 6 node MDS problem considered here is related to a 6 variable, 6 clause Satisfiability problem.

The algorithm used to arrive at a dominating set exploits the fact that any vertex must be either in the set or in the immediate (one edge) vicinity of the set. The problem can then be restated in terms of the neighbourhoods, $N(v)$, for the vertices v of the graph. A dominating set must contain at least one vertex from each $N(v)$. For example, in Figure 2a, neighbourhood $N(p)$ contains b, n, e and p. A dominating set meets the requirements imposed by all six neighbourhoods. Finding just any dominating set is easy, as the original set of all six vertices already contains at least one member of every neighbourhood. Finding the minimal dominating set, however, requires an exhaustive search of all possible dominating sets.

3.2. *Experimental algorithm*

The MDS instance described above can be solved experimentally in two stages: first, generate candidate dominating sets; and second, select the minimal dominating set. All potential solutions were generated from

plasmid pMP6110 using a mix and split methodology (Head et al., 2000), as illustrated in Figure 2b and c. The initial, complete plasmid represents an empty subset. The absence of a station from the plasmid is interpreted as the presence of the corresponding vertex in a subset. All required neighbourhoods are accommodated sequentially. For any neighbourhood, the mixture containing all plasmids is divided in as many test tubes as there are vertices in the neighbourhood. In each of those test tubes, the removal of one specific station is assured by restriction digestion and religation.

After four steps, a library of 16 different plasmids is obtained (Figure 3a). This mixture was transformed to a suitable *E. coli* host strain for protein overexpression (Figure 3b). Protein gel electrophoresis provides neither the resolution nor the sensitivity needed to positively identify proteins. Therefore, the purified protein was analysed using matrix-assisted laser desorption ionisation time-of-flight (MALDI-TOF) mass spectrometry. The resulting spectrum is a representation of all potential minimal dominating sets, with the heaviest protein corresponding to the minimal dominating set (Figure 3c). The original protein (product of plasmid pMP6110) has an average mass of 12054.14 Da. The heaviest protein in the mixture is detected at a mass to charge ratio of 9292. This matches the product of the computational ORF missing stations e and p, which has a predicted average molecular weight of 9291.08 Da. The ORF without these stations in turn corresponds to the minimal dominating set $\{e, p\}$.

4. Discussion

The successful detection of the protein representation of the minimal dominating set shows that mass spectrometry is an attractive read-out strategy for DNA computing. The problem instance solved here is of roughly the same size as molecular computations reported previously (Adleman, 1994; Ouyang et al., 1997; Faulhammer et al., 2000; Liu et al., 2000; Sakamoto et al., 2000). The method is potentially up scalable: MALDI-TOF mass spectrometry is capable of accurately detecting protein mass ranges exceeding 50 kDa (Blank et al., 2002), which would correspond to about 40 plasmid stations and a protein library of 10^{12} species (2^{40}). In large-scale approaches, further information on protein identity can be obtained by application of tandem mass spectrometry techniques (Chalmers and Gaskell, 2000) or proteolytic cleavage of the solutions. This approach is not limited to the plasmid

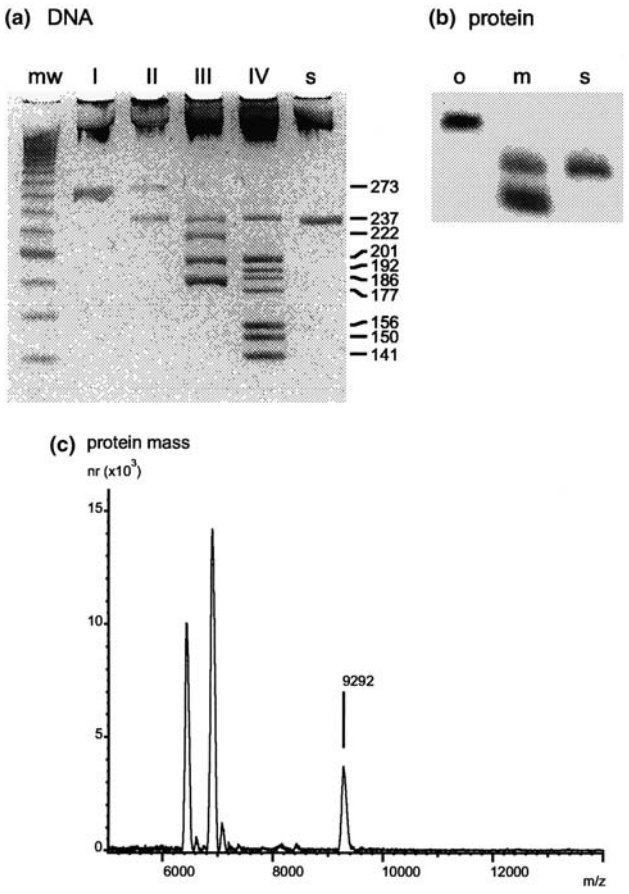

Figure 3. Analysis of potential solutions. (a) 8% polyacrylamide gel with *Eco*RI/*Hin*dIII fragments of religated plasmid after every step (see Figure 2b). Lane mw: DNA size marker; lane s: the isolated solution representing set $\{e, p\}$. (b) Silver stained 12 % SDS-tricine-polyacrylamide gel with purified protein. Lane o: original protein (from plasmid pMP6110); lane m: total protein representation; lane s: isolated solution $\{e, p\}$. (c) MALDI-TOF mass spectrum of the total protein representation. The y axis shows the number of detection events, the x axis the mass to charge ratio of the detected proteins. Since the charge is predominantly +1, this ratio corresponds to molecular weight in Daltons. A single protein species is detected as a mass distribution, because it contains many different isotopes. The average molecular weight of a molecule is determined by allocating the peak of such a spread. Here, the largest protein detected has a molecular weight of 9292 Da.

computing method. If some encoding constraints are taken into account, it is possible to translate answer molecules from any nucleic acid based computation method.

In conclusion, the novel output approach for DNA based computing presented here introduces the use of translation. The ribosome is one of the major information processing components of the cell, and is therefore an interesting candidate as a component of artificial biomolecular computers. So far, only one other design for a hybrid DNA/protein computer has been presented (Sakakibara and Hohsaka, 2003). The generation of protein phenotypes offers possibilities for the implementation of evolutionary algorithms in DNA (Bäck et al., 2000; Chen and Wood, 2000). Protein-based computing methods can also employ the potential of the computational output to function as biologically active molecules. For instance, our plasmid (Figure 1) encodes the plant hormonal peptide *enod40* (Staehelin et al., 2001). In this way, the outcome of a computation could act as a biologically active protein, which in turn switches on downstream computational or biological processes.

Acknowledgements

We thank Pascal van der Wegen, Kees Breek, Ron Hokke and Marco Bladergroen for technical assistance and advice. This work was supported by the Netherlands Organization for Scientific Research (NWO), Exact Sciences.

References

Adleman LM (1994) Molecular computation of solutions to combinatorial problems. Science 266: 1021–1024

Bäck T, Kok JN and Rozenberg G (2003) Evolutionary computation as a paradigm for DNA-based computing. In: Landweber LF and Winfree E (eds) Evolution as Computation (DIMACS Workshop, Princeton, January 1999), pp. 15–40. Springer-Verlag, Heidelberg

Benenson Y, Paz-Elizur T, Adar R, Keinan E, Livneh Z and Shapiro E (2001) Programmable and autonomous computing machine made of biomolecules. Nature 414: 430–434

Blank PS, Sjomeling CM, Backlund PS and Yergey AL (2002) Use of cumulative distribution functions to characterize mass spectra of intact proteins. Journal of the American Society for Mass Spectrometry 13: 40–46

Braich RS, Chelyapov N, Johnson C, Rothemund PWK and Adleman LM (2002) Solution of a 20-variable 3-SAT problem on a DNA computer. Science 296: 499–502

Chalmers MJ and Gaskell SJ (2000) Advances in mass spectrometry for proteome analysis. Current Opinion in Biotechnology 11: 384–390

Chen J and Wood DH (2000) Computation with biomolecules. Proceedings of the National Academy of Sciences of the United States of America 97: 1328–1330

Cox JPL (2001) Long-term data storage in DNA. Trends in Biotechnology 19: 247–250

Faulhammer D, Cukras AR, Lipton RJ and Landweber LF (2000) Molecular computation: RNA solutions to chess problems. Proceedings of the National Academy of Sciences of the United States of America 97: 1385–1389

Garey MR and Johnson DS (1979) Computers and Intractability. A Guide to the Theory of NP-completeness. Freeman, New York

Hagiya M and Ohuchi A (eds) (2003) DNA Computing, 8th International Workshop on DNA Based Computers. Springer-Verlag, Heidelberg

Head T, Rozenberg G, Bladergroen RS, Breek CKD, Lommerse PHM and Spaink HP (2000) Computing with DNA by operating on plasmids. Biosystems 57: 87–93

Head T, Chen X, Nichols MJ, Yamamura M and Gal S (2002) Aqueous solutions of algorithmic problems: emphasizing knights on a 3×3. In: Jonoska N and Seeman NC, (eds) DNA Computing, 7th International Meeting on DNA Based Computers, pp. 191–202. Springer-Verlag, Heidelberg

Jonoska N and Seeman NC (eds) (2002) DNA Computing, 7th International Meeting on DNA Based Computers. Springer-Verlag, Heidelberg

Liu QH, Wang L, Frutos AG, Condon AE, Corn RM and Smith LM (2000) DNA computing on surfaces. Nature 403: 175–179

Mao CD, LaBean TH, Reif JH and Seeman NC (2000) Logical computation using algorithmic self-assembly of DNA triple-crossover molecules. Nature 407: 493–496

Normile D (2002) Molecular computing: DNA-based computer takes aim at genes. Science 295: 951

Ouyang Q, Kaplan PD, Liu S and Libchaber A (1997) DNA solution of the maximal clique problem. Science 278: 446–449

Paun G, Rozenberg G and Salomaa A (1998) DNA Computing. New Computing Paradigms. Springer-Verlag, Heidelberg

Sakakibara Y and Hohsaka T (2003) In vitro translation-based computations. In: Chen J and Reif J (eds) Preliminary Proceedings, 9th International meeting on DNA Based Computers, 1–4 June 2003, Madison, Wisconsin, USA, pp. 175–179. University of Wisconsin, Madison, Wisconsin, USA

Sakamoto K, Gouzu H, Komiya K, Kiga D, Yokoyama S, Yokomori T and Hagiya M (2000) Molecular computation by DNA hairpin formation. Science 288: 1223–1226

Sambrook J and Russell DW (2001) Molecular Cloning: a Laboratory Manual, 3rd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Schagger H and Von Jagow G (1987) Tricine-sodium dodecyl sulfate-polyacrylamide gl electrophoresis for the separation of proteins in the range from 1 to 100 kDa. Analytical Biochemistry 166: 368–379

Staehelin C, Charon C, Boller T, Crespi M and Kondorosi A (2001) Medicago truncatula plants overexpressing the early nodulin gene enod40 exhibit accelerated mycorrhizal colonization and enhanced formation of arbuscules. Proceedings of the National Academy of Sciences of the United States of America 98: 15366–15371

Vieira J and Messing J (1991) New pUC-derived cloning vectors with different selectable markers and DNA replication origins. Gene 100: 189–194