Journal of Combinatorial Optimization, 7, 369–379, 2003

Exordium for DNA Codes*

ARKADII G. D'YACHKOV dyachkov@artist.math.msu.su
Department of Probability, Moscow State University, Moscow

PETER L. ERDÖS elp@renyi.edu.hu
Alfred Rényi Institute of Mathematics, Budapest

ANTHONY J. MACULA macula@geneseo.edu
Mathematics Department, SUNY Geneseo, Geneseo

VYACHESLAV V. RYKOV vrykov@mail.unomaha.edu
Computer Science Department, Northeastern Nebraska University, Omaha

DAVID C. TORNEY dct@lanl.gov
CHANG-SHUNG TUNG cst@lanl.gov
Theoretical Biology and Biophysics, Theoretical Division, Mail Stop K710, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

PAVEL A. VILENKIN paul@vilenkin.dnttm.ru
Department of Probability, Moscow State University, Moscow

P. SCOTT WHITE scott_white@lanl.gov
Genomics, Biosciences Division, Los Alamos National Laboratory

Received February 5, 2003; Accepted July 29, 2003

Abstract. We describe how deletion-correcting codes may be enhanced to yield codes with double-strand DNA-sequence codewords. This enhancement involves abstractions of the pertinent aspects of DNA; it nevertheless ensures specificity of binding for all pairs of single strands derived from its codewords—the key desideratum of DNA codes– i.e. with binding feasible only between reverse complementary strands. We defer discussing the combinatorial-optimization superincumbencies of code construction. Generalization of deletion similarity to an optimal sequence-alignment score could readily effect advantageous improvements (Kaderali, Master's Thesis, Informatics, U. Köln, 2001) but would render the combinatorics opaque. We mention motivating applications of DNA codes.

Keywords: binding propensity, DNA computing, DNA hybridization, digital velcro, formulation, directed design

*LAUR no. 02-5807.

1. Pertinent aspects of DNA

Single strands of DNA are, abstractly, (A,C,G,T)-quaternary sequences, with the four letters denoting the respective nucleic acids. Strands of DNA sequence are oriented; thus, AACG is distinct from GCAA. Furthermore, in nature DNA is ordinarily *double stranded*: each sequence, or strand, occurs with its *reverse complement*, with reversal denoting that the sequences of the two strands are oppositely oriented, relative to one other, and with complementarity denoting that the allowed pairings of letters, opposing one another on the two strands, are {A, T} or {C, G}—the canonical Watson-Crick pairings (Bell and Torney, 1993). Reverse complementation also occurs in pure mathematics (Johnson, 1997, p. 4).

Therefore, to obtain the reverse complement of a strand of DNA (i) reverse the order of the letters and (ii) substitute each letter with its complement. For example, the reverse complement of AACGTG is CACGTT. A double strand results from adjoining reverse complementary strands in opposite orientations:

As a mnemonic, the single strand oriented left to right is orthodoxly penned, whereas the oppositely oriented strand is sinistrally penned—upside-down and backward. Our esotropic depiction, paraphrasing conventional notation, emphasizes the through-the-looking-glass aspects of DNA: the interchangeability of its strands.

In this convoluted domain, just as tweedledum evokes tweedledee, a strand evokes its reverse complement, and our DNA codes are composed of double strands whose individual strands are reverse complementary. A measure of similarity for pairs of codewords of DNA codes should, in a nutshell, model the favorability of pairing between (the four pairs of) non-reverse complementary strands because only the reverse-complementary pairs of strands should bind to one another. Coding theory, remarkably, very nearly anticipated these particulars.

It may be noted that DNA also evokes a novel poset of finite, quaternary sequences with an ordering given by the inclusion of one sequence or its reverse complement as a subsequence of another. This poset is currently under consideration, making analogy to the classical *word poset* (which omits the reverse-complement inclusion) (Erdös et al., in prepration; Erdös, Torney and Sziklai, 2001).

2. Binding specificity within DNA codes

Herein, we toe a discrete mathematical line, i.e. the present note establishes an abstraction, formulating a DNA code suited to our desiderata. An allegory involving thermodynamics will serve to introduce the notion of *binding specificity*.

Consider the following thought experiment. Separate the single strands of the codewords and, then, mix them all back together again, letting them find their own way to minimum energy by producing aggregates of strands. Specificity of binding connotes that, at this equilibrium, the double-strand codewords will all be present, adjoined as depicted above, and that, effectively, no other aggregates (including singlets and doublets) will occur. This

fastidiousness of binding is the *raison d'être* for DNA codes, as corroborated by instances of their applications (viz Section 7).

3. Combinatorics of binding specificity

A natural abstraction of binding specificity is to base it upon the maximum number of Watson-Crick bonds (complementary letter pairs) which may be formed between two oppositely oriented strands. For two reverse complementary strands, this number plainly equals their length, but to venture beyond this requires gumptious assumptions.

Were DNA strands inflexible, then Hamming codes would be appropriate (Marathe, Condon and Corn, 2001). For strands of length ten or greater, it is more circumspect to try the opposite tack, considering strands to be fully flexible—"rubber-band" DNA—yielding the following *ansatz*.

The binding propensity of two single strands is, to a first approximation, measured by the length σ of the longest common subsequence (not necessarily contiguous) of either strand and the reverse complement of the other.

σ plainly doesn't depend upon which strand is selected (the unselected strand is reverse complemented), and this combinatorial similarity measure abstracts DNA-sequence binding propensity. Through free-energy minimization, kindred subsequences are assumed to bind to one another, and letters remaining unpaired are assumed to be peripheralized by means of loops. After reverse complementation of one sequence, kindred sequences are subsequences in common.

Given the strand and the reverse-complement strand, optimal sequence alignment thereof is the method of choice for determining σ (Needleman and Wunsch, 1970). This procedure may be viewed as apposite insertion of loops in the two sequences, as needed, to maximize the resulting number of identical aligned letters. The "dynamic programming" algorithm is optimal for computing σ: with complexity equal the product of the lengths of the two strands (Smith and Waterman,1981).

Example 1. AACGTG and its reverse complement CACGTT have $\sigma = 4$ thus, either may bind to itself.

For some purposes, e.g., viz figures 3 and 4, more detailed modeling of the thermodynamics of DNA binding may be desirable, mitigating the idealizations of DNA strands as fully rigid or as fully flexible. Optimal sequence alignment may also be used to achieve this aim (Kaderali, 2001). For example, the different binding energies of C-G and A-T pairs could be modeled by assigning an alignment score of, say, 1.9 to C and G identities and by assigning a score of 1.0 to A and T identities (recalling that these letters are derived from one sequence and the reverse complement of the other, as mentioned above). Mismatches and loops could be taken to accrue a score of zero. Then, the optimal sequence alignment would determine an optimal disposition of letter pairings and loops which maximizes the sum of the scores of the paired letters.

Below its *melting temperature* the preferred configuration of two strands, reverse complementary or no, is their aggregate. Above this temperature two single strands are preferred. Nearest-neighbor models constitute the state of the art for predicting melting temperature (Breslauer et al., 1986) This model embodies the dependencies of the binding energy for a given pair of letters upon the neighboring pair. Sequence alignments have been adapted to comprise both nearest-neighbor and single-strand-loop models, the latter for unpaired letters (Kaderali, 2001). Bounding a well-chosen optimal sequence alignment score pegged to the nearest-neighbor melting temperatures—for codewords and for the untoward binding of other pairs of codeword strands—could, someday, yield a superior DNA code (Kaderali et al., 2003).

The present aim, however, is to clearly define an elemental DNA code, based upon a maximum allowed length of a common subsequence for strands derived from distinct codewords. Our panoptic formulation of DNA binding, based on restricting σ, includes both feasible and infeasible configurations—the latter ruled out by the physical properties of DNA. By taking a generous view of what could "go wrong", untoward binding will plainly be prohibited; the cost for also restricting "what can never be" is that "what can be" will escape maximum correction.

4. Coding-theoretic background

For two q-ary n-sequences $\mathbf{a}$ and $\mathbf{b}$, the longest length of a sequence occurring as a (not necessarily contiguous) subsequence of both is called a *deletion similarity* between $\mathbf{a}$ and $\mathbf{b}$. It may be noted that n minus the deletion similarity is a metric called the *deletion distance* (cf. Hollman, 1993; Levenshtein, 1966).

Example 2. Let $q = 2$, $n = 8$, $\mathbf{a} = (0, 1, 0, 1, 1, 0, 1, 1)$ and $\mathbf{b} = (0, 0, 1, 0, 0, 1, 0, 1)$. The conventional Hamming similarity (i.e., the number of shared, aligned digits) between $\mathbf{a}$ and $\mathbf{b}$ is equal to 2. The deletion similarity between $\mathbf{a}$ and $\mathbf{b}$ is equal to 6 because 6-sequence $(0, 0, 1, 0, 1, 1)$ is as long as any sequence occurring as a subsequence in both $\mathbf{a}$ and $\mathbf{b}$. Two other examples of common subsequences of maximum length are $(0, 1, 0, 1, 0, 1)$ and $(0, 0, 1, 1, 0, 1)$.

Codes of length n with an upper bound σ, $0 \leq \sigma \leq n - 1$, on the deletion similarity between any pair of codewords $\mathbf{a}$ and $\mathbf{b}$ are called deletion-correcting codes (Levenshtein, 1966; Ullman, 1967). They are, in fact, capable of correcting any combination of $\leq n-1-\sigma$ deletions and insertions.

Deletion-correcting codes were entertained, beginning in the 1960s, in connection with correction of synchronization errors (Levenshtein, 1966; Ullman, 1967). Coding theorists refer to such codes as *directed packings* (cf. Colbourn and Dinitz, 1996, Section 15). We have the following asymptotic ($n \to \infty$) upper bound on code size for binary codes with codewords of length n (Levenshtein, 1966):

$$\frac{2^n(n-1-\sigma)!}{n^{n-1-\sigma}};\ 0 \leq \sigma \leq n-1.$$

With the exception of the cases $\sigma = n - 2$ (Varšamov and Tenengol'ts, 1965) and $\sigma = 2$ (Yin, 2001) (and some special cases (Shalaby, Wang and Yin, 2002)), no general, deterministic constructions are known for deletion-correcting codes, at odds with the cornucopia thereof for Hamming codes.

5. DNA code definition

Definition 1. A DNA code is a set of (A,C,G,T)-quaternary n-sequences satisfying

(1) Codewords are double-strands (composed of reverse complementary strands). Hence, the reverse complement of each strand in the code also occurs in the code.
(2) No strand in the code equals its reverse complement.
(3) The deletion similarity of distinct strands is less than or equal to σ, with $0 \leq \sigma < n$.

The third condition, by itself, specifies a deletion-correcting code. A sequela of conditions one and three is that (even-length) subsequences invariant to reverse complementation and of length exceeding σ are forbidden to occur as subsequences of codeword strands.[1]

For sufficiently small σ, these conditions are seen to engender the desired specificities of binding; for didactic purposes, consider any bipartition of the codeword strands such that each reverse-complementary pair is disunited across its two blocks. (For some applications, as will be seen in Section 7, it is advantageous to create two such blocks). Recall that the condition on σ diminishes the binding propensity for either of the codeword strands with the reverse complement of the other. Therefore, the implications of the third condition are (a), from application within blocks, no inter-block binding—excepting, of course, reverse-complementary pairs of codeword strands, as depicted in figure 1(a)—and (b), from application to all pairs of codeword strands with strands derived from both blocks, no intra-block binding—including the binding of a codeword strand with itself, as shown in figure 1(b). Together, these plainly comprise all pairs of codeword strands, and, as sought, the only feasible binding is between the reverse-complementary strands of codewords. Furthermore, included in the prospective binding configurations are all the alignments with no loops entering into codes based on the Hamming distance.

Note that $10 \leq n \leq 40$ is experimentally accessible and that codes with more than 10^9 codewords could soon be called for. In this context, it is noteworthy that it is possible to construct hexanary and octanary DNA codes, using additional, synthetic pairs of nucleic acids, for example, iso-C and iso-G, which bind only with each other.

6. Preliminary results on DNA codes

Some of the novel aspects of DNA codes were first explored in the context of conventional, Hamming-like codes because these are much simpler (Marathe, Condon and Corn, 2001). We considered the ramifications of using different similarities for matching A and matching T versus those for matching C and matching G (D'yachkov and Torney, 2000). The consequences of imposing the first two conditions upon a Hamming code were explored with linear codes constructed from invertible cyclic codes (Rykov et al., 2000).

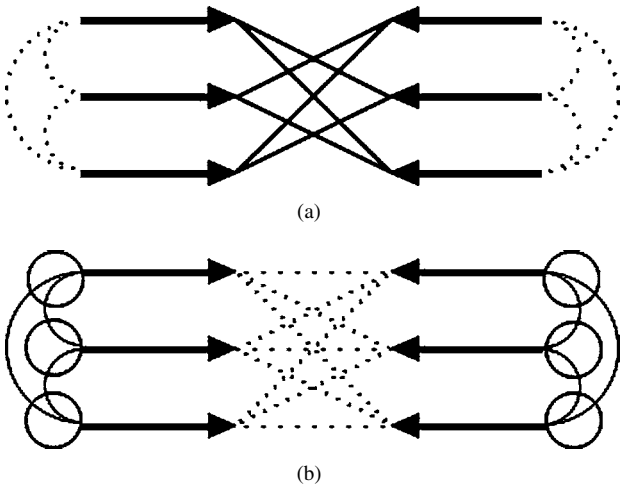

Figure 1. Interdictions on binding. Here a DNA code with three words is depicted, and their six strands have been partitioned into two blocks of three sequences, with one strand from each codeword in each block. This is the signification of the six thick horizontal arrows, with the right-hand and left-hand sides constituting the two blocks and reverse-complementary sequences appearing adjacent to one another. In figure 1(a), the restrictions on σ for the pairs of sequences within the blocks are indicated by the dashed arcs. Their non-binding implications are indicated by the (central) six solid lines between the codeword strands. Thus, the only allowed binding is between reverse complementary strands constituting the codewords. In figure 1(b), the restrictions on σ for pairs of strands with one strand from each of the two blocks are indicated by the dashed lines, engendering a complete bipartite graph. Their non-binding implications are indicated by twelve solid arcs: six duplicating those of figure 1(a) and six new loops, connoting the non-binding of the code's strands with themselves. These depictions obviously generalize to DNA codes of arbitrary size. Horizontal dashed lines indicate interdiction of homo-dimer formation (i.e., self-binding of strands).

We have also obtained a random-coding lower bound on DNA-code size, which may be stated as follows (D'yachkov et al., submitted).

Theorem 1. *As* $n \to \infty$, *DNA code size grows exponentially with* σ, *provided that* $0.73n \leq \sigma$.

It may be noted that 0.73 is an approximation for a root of a transcendental equation.

Remark 1. We predict that DNA code size in fact grows exponentially with n, provided that $ns_{\text{dna}} \leq \sigma$, where the constant s_{dna}, $0 < s_{\text{dna}} < 0.73$, may be derived from the average asymptotic ($n \to \infty$) ratio of the deletion similarity to n between quaternary n-sequences and their reverse complements in the space of all 4^n quaternary n-sequences. Using the method of least squares (and a special implementation of the Monte Carlo method), we constructed (and calculated) point estimators and confidence intervals for s_{dna} (D'yachkov et al., submitted). For instance, the 95%-confidence interval for s_{dna} is $0.6025 \leq s_{\text{dna}} \leq 0.6035$.

Thus, for instance, it should be feasible to construct a large-size DNA code with the parameters $\sigma = 15$ and $n = 20$: parameters that should also yield sufficient binding

specificity for many applications. We have explored heuristic methods for computational construction of random DNA codes (D'yachkov et al., submitted).

7. Applications of DNA codes

DNA codes have many potential applications—as signified by following two examples.

7.1. Digital Velcro

The moniker for DNA codes Digital Velcro is well deserved, with "hooks" and "loops" being its single strands. DNA codes will be essential for the implementation of large-scale biological experiments in parallel and in a small volume: for instance, the efficient determination of specific variants of 10^5 genes occurring in an individual's genome: a mantra of molecular medicine.

For illustration, experiments may be performed, using single strands from a DNA code to tag the results. Then, to facilitate readout, the latter may be "arrayed" using the reverse-complementary strands, as illustrated in figure 2. Note how intra-block binding of strands (cf. Section 5) would compete with the desired, double-strand formation.

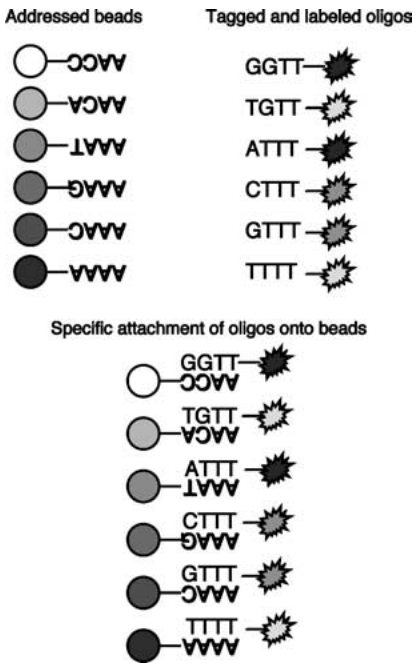

Figure 2. Digital Velcro. In this figure a fanciful DNA code of six codewords of length four is depicted (one whose σ was not determined). Strands are referred to as "oligos". Imagine that experiments (castaneaceous objects on the right) are tagged with the strands from one block of strands, and that these are to be affixed to beads of different colors (circular objects on the left) using the other block, containing the reverse complementary strands. In fact, 10^5 or more experiments could be performed and interrogated in parallel, using a suitably sized DNA code.

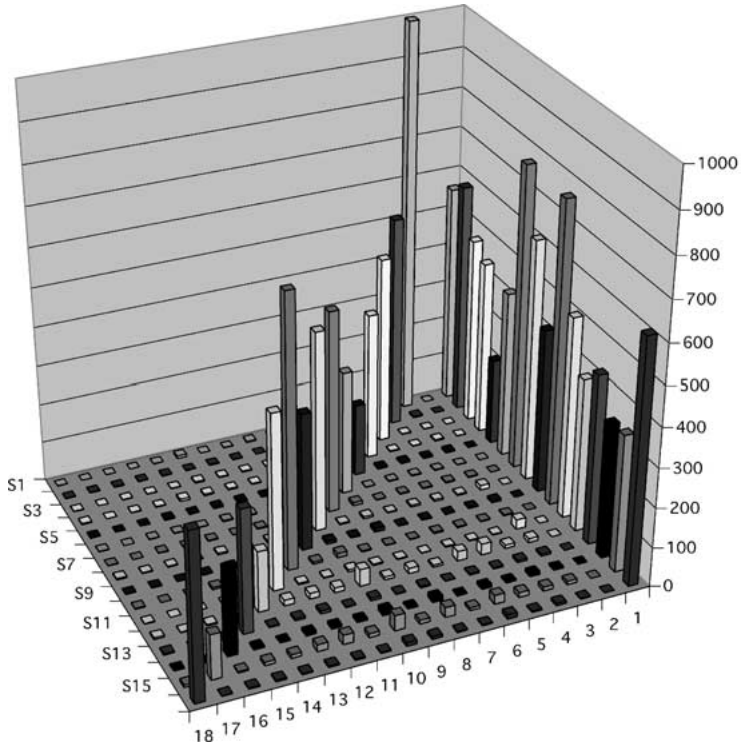

Figure 3. *Binding in a deletion-correcting code*. This figure illustrates, by the heights of the histograms, the result of an experiment in which the concentration of bound strand pairs for each of 16 strands of a random, deletion-correcting code (of 256 codewords with $\sigma = 13$ and $n = 20$ (Cai et al., 2000) and the reverse complements of these strands was measured. Thus, in all, the results of 256 experiments are depicted. (In addition, rows 1 and 2 contain the results from control experiments, which may be ignored for the present purposes.) The heights of the bars indicate these concentrations. It is observed that the binding between strands and their reverse complement is dominant, although inopportune binding also occurs at lower levels. (Generalization of deletion similarity to optimal sequence alignment scores has the potential of yielding greater fastidiousness of binding).

We constructed a small, random, deletion-correcting code with $n = 20$, $\sigma = 13$ and code size of 256 (Cai et al., 2000). The binding of 16 of these codewords strands with the reverse complements thereof is illustrated in figure 3. Binding between two of these codeword strands—not prohibited in this construction—is depicted in figure 4.

7.2. DNA computing

It is testament to the naturalness of our definition of DNA codes that they are also wetware for the current prototyping of DNA computing for combinatorial problems (Adleman,1994; Ouyang et al., 1997; Sakamoto et al., 2000). In each of these experiments, the computation hinges upon specificity of binding between DNA sequences and their reverse complements. The binding of a sequence to its reverse complement is, in fact, a "gate" in DNA computers,

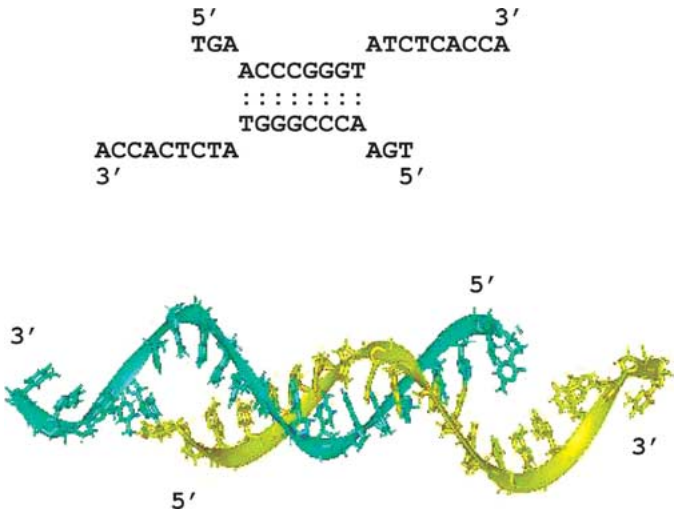

Figure 4. *Binding between strands of a deletion-correcting code*. Recall that a deletion-correcting code omits the first two conditions satisfied by a DNA code, and, hence, binding between its codewords, implemented as single strands, has not been prevented. The deletion-correcting code attempts only to ensure that each strand will bind selectively with its reverse complement, out of all the reverse complements. Using our insights into DNA structures, we identified a plausible binding between two strands of this code. This pair of strands is depicted: both abstractly and also with a projection of a three-dimensional model, illustrating the nucleotides' shapes and positioning. Ribbons denote the sugar "backbones". The abstraction denotes nucleotide-nucleotide binding by colons. Note that 5′ to 3′ is the the standard molecular-biological terminology for the canonical "left-right" DNA strand. The letters on the other strand, having the opposite orientation, have been written according to conventional notation; please forgive this lapse into orthodoxy. We advocate the depiction of this strand by ʇƃɐʇƃɐɐɔɔɔƃƃʇɐʇɔʇɔɐɔɔɐ.

and uncontrolled errors would occur were sequences capable of binding to other than their reverse complement.

Consider, for instance, Adleman's pioneering computation of Hamiltonian paths in a digraph. Vertices are assigned DNA sequences and the edges are assigned the catenation of the reverse complements of the prefix half and suffix half of the respective vertex sequences. Thus, the digraph is replicated through the selection of two codewords from a DNA code—and specifying which strands thereof to catenate together—to constitute a sequence for each vertex. Only were there specificity of binding would a long double strand result from this imbrication.[2] Upon reflection, a specialized DNA code could offer further advantages for this application because, for instance, the reverse-complement of a suffix needs to bind specifically to the suffix, even when the suffix is attached to a prefix. As DNA codes have not yet been applied in this context, however, such refinements are as yet unwarrantable.

8. Summary

Although we have achieved modest characterization of suitable parameter choices for large-size codes, we lack algebraic methods for code construction, and it remains a challenge to

determine various bounds on DNA-code size. Methods for constructing random codes need to be evaluated and implemented. One can imagine additional measures of sequence similarity which could more effectively restrict attention to the domain of feasible binding configurations, but the necessity of pursuing these is not yet clear. The sledge would involve optimal, pairwise sequence alignment scores using a raft of parameters (Kaderali, 2001; Needleman and Wunsch, 1970.)

We have attempted deconvolution; the reader may judge our success. Were the foregoing to contain any lore, its selvage would herald that advancing the state of the art on all fronts, is essential, in part, because every advance enables future applications. Thus, "do all you know and try all you don't..." (L. Carroll, *The Hunting of the Snark*).

Acknowledgments

This manuscript is dedicated to the memories of Drs. George I. Bell and Walter B. Goad, pioneers in bioinformatics and founders of the Theoretical Biology Group at Los Alamos National Laboratory. A first pass at clarifying the content of this note was afforded by the meeting on Emerging Applications of Combinatorial Designs, held at the MSRI (UC Berkeley) in November, 2000. The ambiance of SETA '01 spurred this manuscript on to completion. We are indebted to Professors F. Hwang and D.-Z. Du for including our manuscript in these proceedings. This work was supported by the USDOE, under Contract W-7405-ENG-36. P.L.E. was partially supported by the Hungarian NSF under no. T29255 and no. T34702 and also by the A.V. Humboldt Foundation. D.C.T. and P.S.W. were partially supported by internal funding: LDRD#X1FC and XAPN (to P.S.W.).

Notes

1. We are indebted to Professor Aiden A. Bruen, of the University of Calgary, for this observation.
2. Detailed inspection of double-strands of appropriate length is used to validate that all vertices are represented once (Adleman,1994).

References

L.M. Adleman, "Molecular computation of solutions to combinatorial problems," *Science*, vol. 266, pp. 1021–1024, 1994.

G.I. Bell and D.C. Torney, "Repetitive DNA sequences: Some considerations for simple sequence repeats," *Computers and Chemistry*. vol. 17, pp. 185–190, 1993.

K.J. Breslauer, R. Frank, H. Blocker, and L.A. Markey, "Predicting Duplex DNA Stability from the base sequence," *PNAS USA*, vol. 83, pp. 3746–3750, 1986.

H. Cai, P.S. White, D.C. Torney, A. Deshpande, Z. Wang, B. Marrone, and J.P. Nolan, "Flow cytometry-based minisequencing: A new platform for high throughput single nucleotide polymorphism scoring," *Genomics*, vol. 66, pp. 135–143, 2000.

C.J. Colbourn and J.H. Dinitz, *The CRC Handbook of Combinatorial Designs*. CRC Press: Boca Raton, 1996.

A.G. D'yachkov and D.C. Torney, "On similarity codes," *IEEE Trans. on Information Theory*, vol. 46, pp. 1558–1564, 2000.

A.G. D'yachkov, D.C. Torney, P.A. Vilenkin, and P.S. White, "Reverse—Complement similarity codes," *IEEE Trans. on Information Theory*, submitted.

P.L. Erdös, P. Ligeti, P. Sziklai, and D. C. Torney, "Subwords in reverse complement order," in preparation.
P.L. Erdös, D.C. Torney, and P. Sziklai, "A finite word poset," *Elec. J. of Combinatorics*, vol. 8, 2001.
H.D.L. Hollman, "A relation between Levenshtein-type distances and insertion and deletion correcting capabilities of codes," *IEEE Trans. on Information Theory*, vol. 39, pp. 1424–1427, 1993.
D.L. Johnson, *Presentation of Groups*, 2nd edition. London Mathematical Society Student Texts, Cambridge University Press: Cambridge, vol. 15, 1997.
L. Kaderali, "Selecting target specific probes for DNA arrays," Master's Thesis, Informatics, U. Köln, 2001.
L. Kaderali, A. Deshpande, J.P. Nolan, and P.S. White, "Primer-design for multiplexed genotyping," *Nucleic Acids Research*, vol. 31, pp. 1796–1802, 2003.
V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *J. Soviet Phys. Doklady*, vol. 10, pp. 707–710, 1966.
A. Marathe, A.E. Condon, and R.M. Corn, "On combinatorial DNA design," *J. Comp. Biol.*, vol. 8, pp. 201–219, 2001.
S.B. Needleman and C.D. Wunsch, "A general method applicable to the search for similarities in the amino-acid sequences of two proteins," *J. Mol. Biol.*, vol. 48, pp. 443–453, 1970.
Q. Ouyang, P.D. Kaplan, S. Liu, and A. Libchaber, "DNA solution of the maximal clique problem," *Science*, vol. 278, pp. 446–449, 1997.
V.V. Rykov, A.J. Macula, C.M. Korzelius, D.C. Englehart, D.C. Torney, and P.S. White, "DNA sequences constructed on the basis of quaternary cyclic codes," in *Proceedings of the 4th World Multiconference on Systematics, Cybernetics, and Informatics, SCI 2000/ISAS 2000*, Orlando, Florida, July 2000.
K. Sakamoto, H. Gouzu, K. Komiya, D. Kiga, S. Yokoyama, T. Yokomori, and M. Hagiya, "Molecular computation by DNA hairpin formation," *Science*, vol. 288, pp. 1223–1226, 2000.
N. Shalaby, J.M. Wang, and J.X. Yin, "Existence of perfect 4-deletion-correcting-codes with length six," *Designs, Codes and Cryptography*, vol. 27, pp. 145–156, 2002.
T.F. Smith and M.S. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, vol. 147, pp. 195–197, 1981.
J.D. Ullman, "On the capabilities of codes to correct synchronization errors," *IEEE IT*, vol. 13, pp. 95–105, 1967.
R.R. Varšamov and G.M. Tenengol'ts, "One-asymmetrical-error correction codes (in Russian)," *Avtomatika I Telemekhanika*, vol. 26, pp. 288–292, 1965.
J. Yin, "A combinatorial construction for perfect deletion-correcting codes," *Designs, Codes, and Cryptography*, vol. 26, pp. 99–110, 2001.