Computationally Inspired Biotechnologies:

Improved DNA Synthesis and Associative Search Using Error-Correcting Codes & Vector-Quantization

John H. Reif and Thom LaBean

Computer Science Department

Duke University

Adleman: "Computationally Inspired Molecular Technology"

- **Molecular Technology that is inspired by computational methods.**
  - **A visionary concept which seems destined to have major impact.**

Computationally Inspired Biotechnologies

- Biotechnology methods that are inspired by computational methods.
  - Take inspiration from methods used in computer science and related disciplines.
  - Apply these to develop improved biotechnology.
- Example:
  - PCR operates by a recursive doubling of concentrations of selected DNA strands.
  - The inventor of PCR has stated he was inspired by the technique of recursive programming in his discovery of PCR.
- Challenge:
  - CS techniques need to be re-tailored to work in the biotechnology context.

Computationally Inspired Biotechnologies: Applying Coding Theory

Adapt information theoretic techniques which originating in computational and information processing disciplines:

- **Error-Correcting Codes:**
  - Developed to correct transmission errors in electronic media.
  - We adapt to:
    - decrease (in certain contexts, optimally) error rates in optically-addressed DNA synthesis (e.g., of DNA chips).

- **Vector-Quantization (VQ) Coding techniques:**
  - Developed to correct transmission errors in electronic media.
  - Adapt to decrease (in certain contexts, optimally) error rates in optically-addressed DNA synthesis (e.g., of DNA chips).
  - Used to cluster, quantize, and compress data such as speech and images.
  - We adapt to Improve:
    - I/O rates (in certain contexts, optimally) for transformation of electronic data to and from DNA with bounded error.
    - Associative search in DNA databases by reducing the problem to that of exact affinity separation.

Biomolecular Computing (BMC): Using of biotechnology to do computation.

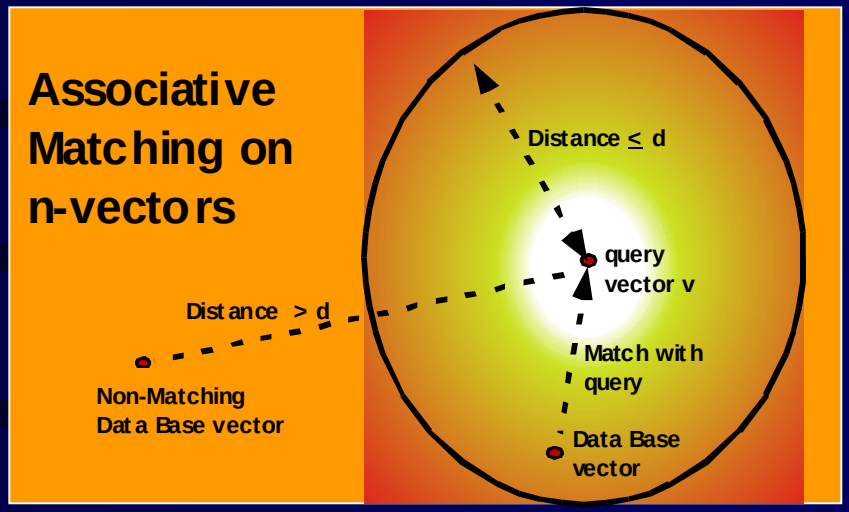
- BMC has impressive potential for molecular-scale computation.
- Recombinant DNA technology operates on vast numbers of DNA strands in massively parallel fashion.
- Ultra-Compact DNA Storage Media:
  - Very large amounts of data that can be stored in compact volume.
  - Vastly exceeds the storage capacities of conventional electronic, magnetic, or even optical media.
  - DNA is about 10^8 times more compact than other storage media currently being used.
  - A gram of DNA contains about 10^{21} DNA bases = about 10^8 terabytes.
  - A few tens of grams of DNA may have the potential of storing all the human-made data currently stored in the world.
  - Most recombinant DNA techniques can be applied at concentrations of about 5 grams of DNA per liter of water.

Killer Applications for BMC.

- **Challenge:**
  - Find applications of BMC that have:
  - Commercial utility in the near term.
  - Resource requirements (number of recombinant DNA steps, volume of test tubes, etc.) should scale well so future large scale demonstrations will be feasible.

- **BMC potential applications. Demonstrations and Proposals:**
- **SAT Problems:**
  - Solution of small size combinatorial search problems using Separation techniques and surfaced based chemistry
    - Not scalable (volume grows exponentially with problem size)
    - But can do useful massively parallel Boolean processing !
- **Scalable applications:**
  - Methods for hiding DNA and for encrypting DNA
  - Neural network learning
  - Nano-assemblies
    - For Nano-Electronics: Placement via DNA associative matching
    - For surface attachments: Ultra-scale DNA chips

Example Application of BMC: Associative Search



- ***Database***:
  - Ordered list of elements of n-vectors, whose elements range over a finite range.
  - Each vector of database has a unique identifying *index* in the database.
- Associative Search Query:
  - Query vector v
  - Distance bound d.
    - distance(u,v) = $|u_1-v_1|+ |u_2-v_2|+\ldots |u_n-v_n|$
- Find *distance d near-matches:*
  - Search the entire database for those vectors of the database that are of distance at most d from the query vector.
- *Closest match:*
  - Find the index to a vector of the database of smallest distance from the query vector.

Example:
Associative Search in Image Database

- **Preprocessed by a procedure A forming an *attribute database*:**
  - **List of low level image attributes for each image or sub-image.**
- **Given an input image I,**
  - **Use A to determine its vector A(I) of image attributes.**
- **Associative search in the attribute database provides the closest match to A(I).**
  - **Provides an index to that image in the image database whose attributes best match that of the input image I.**

DNA Annealing as a Massively Parallel Associative Search Engine.

- Ultra-compact DNA storage media.
  - The DNA databases might have size n = 1000 terabytes.
- Supports highly parallel associative searches within the entire media.
  - [Baum]: proposed using known recombinant DNA methods for DNA ligation affinity separation. Possible methods:
    - Streptavidin-coated paramagnetic beads.
    - PCR uses DNA annealing to amplifying the frequency of those DNA strands that have a particular chosen sequence.
  - Use of DNA words for vector elements:
    - Each element of a vector of the database is encoded by a DNA word.
    - Each n-vector v of database encoded by sequence of n DNA words, followed by DNA word for identifying index to v.
- Scalable:
  - If < maximum concentration, # of recombinant DNA operations and volume are independent of database size.
  - Otherwise, # of recombinant DNA operations and volume linear in the size of the database.
- Many remaining issues were not considered by Baum.

DNA Annealing as a Massively Parallel Associative Search Engine.

Major Challenges Remaining:

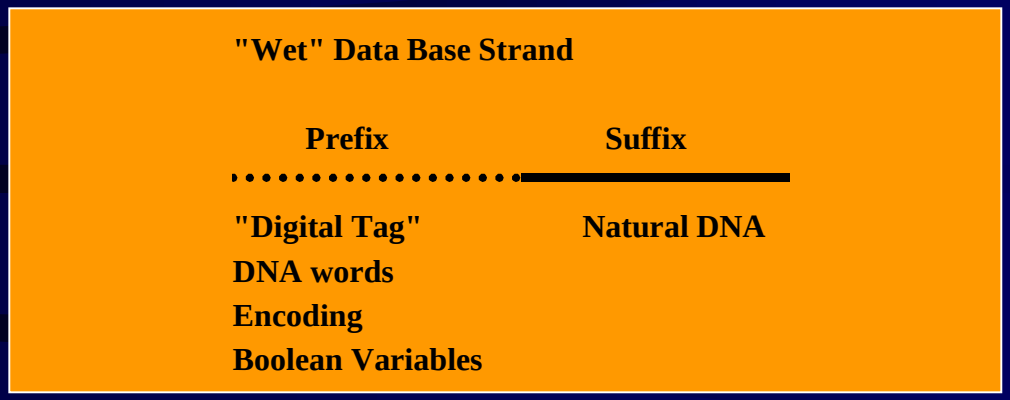
(not considered by Baum)

- **(a) Input and Output (I/O) to Conventional Media:**
  - Goals: error-resiliency and optimal I/O rate for a given error rate.
- **(b) Refining the Associative Search to Exact Affinity Separation:**
  - The query may not be an exact match or even partial match with any data in the database.
  - DNA annealing affinity methods:
    - Work best annealing on complementary sequences.
    - Do not perform well for associative matching in case of partial matches with scattered mismatches in interior of vectors.
- **(c) Extension to Include Boolean Conditionals:**
  - Extend associative search queries to Boolean formula conditionals (with a bounded number of Boolean variables), by combining our methods for DNA associative search with known BMC methods for solving the SAT problem.
  - Example: extended queries executed on:
    - Natural DNA strands (from blood or other tissues)
    - Appended with DNA words encoding binary information about each strand (e.g, the social security number of the person whose DNA was sampled, cell type, the date, further medical data, etc.).

Motivating Example
Improved biotechnology methods:
Massively parallel associative search in extremely large databases encoded as DNA strands.

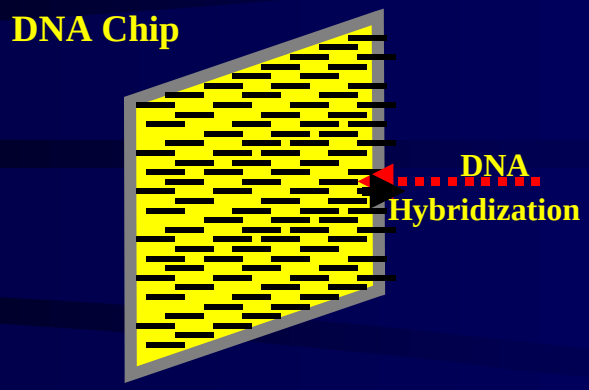
- **We use improved biotechnology techniques based on Error-Correction and VQ Coding.**
  - **The database may initially be in conventional (electronic, magnetic, or optical) media, rather than the form of DNA strands.**
    - **Proposed Solution: Apply DNA chip technology improved by Error-Correction and VQ Coding methods for error-correction and compression.**
  - **The query may not be an exact match or even partial match with any data in the database, but DNA annealing affinity methods work best for these cases.**
    - **Proposed Solution: Apply various VQ Coding methods for refining the associative search to exact matches.**
  - **Extend associative search queries in DNA databases to include Boolean formula conditionals (with bounded # of Boolean variables).**
    - **Proposed Solution: Combine our methods for DNA associative search with known BMC methods for solving small size SAT problems.**

Digital Tagged Natural DNA



- DNA strands augmented with prefix "digital tag strands" consisting of a sequence of DNA words encoding Boolean values.
- Example:
  - The extended database might consist of natural DNA strands (e.g., from blood or other body tissues)
  - Appended "digital tag strands" consisting of DNA words encoding identifying information about each strand (such as social security number of the person whose DNA was sampled, cell type, the date, further medical data, etc.).
- The "digital tag strands" may have been constructed by previous BMC processing.

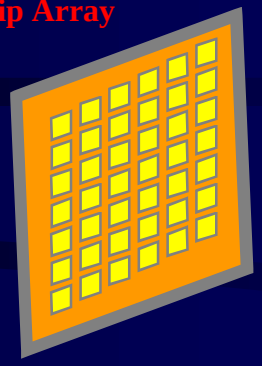
DNA Chips



- Individual DNA chips give highly parallel input/output over 2D surfaces.
- Use for Input:
  - use of photosensitive DNA-on-a-chip technology:
    - 2D optical input is converted to DNA strands encoding the input data.
- Use for Output:
  - Via hybridization at the sites with fluorescent labeled DNA, the output can be read as a 2D image.
- Scaling of Individual DNA chips:
  - Each DNA chip can be optically addressed at up to 10^5 sites.
  - Projected to be millions of sites in the immediate future.

Massively Parallel I/O Using Arrays of DNA Chips

DNA Chip Array

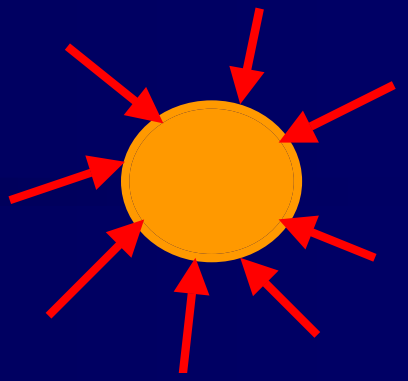


- I/O between:
  - conventional electronic media
  - and a "wet" database of DNA strands
    - (in solution or on solid support).
- Propose Solution: Large Arrays of DNA Chips:
  - A few thousand chips can be placed on a 2D array compact enough so all chips can be addressed by a single optical system.
  - Gives potential of parallel synthesis of DNA at 10^8 sites or more to many billions of sites.
- Massively parallel DNA input/output:
  - Has a potential for achieving a rate of I/O to convention optical/electronic media in the order of gigabit rates or more.

Massively Parallel I/O using Arrays of DNA Chips
Technical Challenge:
Error rates due to optically addressed base synthesis.

- Most common error in optically addressed synthesis of DNA is a premature truncation and deletion in the growing strand.
  - Error rate in optically addressed DNA synthesis methods used for DNA chips is roughly 4% to 8% per base
  - Corresponds to an expected error in every 12 to 25 base pairs.
- Application of DNA chips for I/O in BMC seems limited by current error rates.
  - Each DNA strand synthesized may be quite long (over 25 bases per strand).
  - Majority of DNA strands expect > one synthesis error.
- Commercial DNA chips (proprietary Affymetrix technology)
  - Synthesis error rates not known for each type for possible error.
  - Utilize only a fraction of 10^5 optically addressable sites.
    - Current maximum: about 42,000 sites
    - Typical DNA chip: uses about 7,000 sites
  - Today: currently synthesis error rate seems not the dominant limiting factor,
  - Future: will impact scalability (addressable sites & strand length) of DNA chip technology.

3D Fluid Micro-array Technology using Beads



- DNA solid supported to Beads
- Can use fluorescence tags
- Bead sizes: 3 to 100 microns
- Bead material: plastic or polystyrene

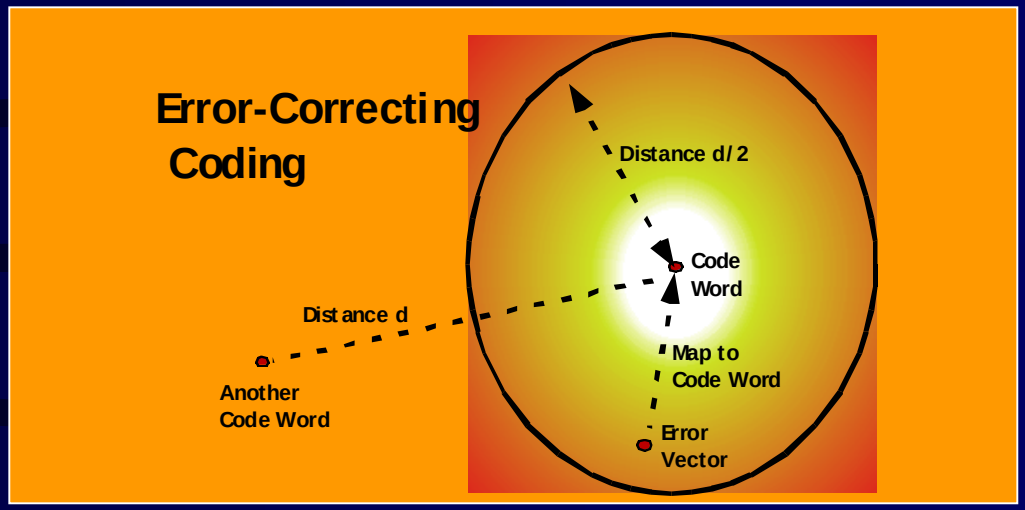
-Readout Methods for Beads:

- Fluorescence activated cell sorter (FACS)
  - Example: MoFlo cell sorter
- Fiber Optic Readout: Illuminata, Inc.
  - 60,000 fibers each of 3.5 microns
  - Etch ends of fibers and then add attachment chemistry to attach a bead to each fiber.

-Combinatorial Libraries of Digital Tags appended to Natural DNA or RNA: Lynx, Inc

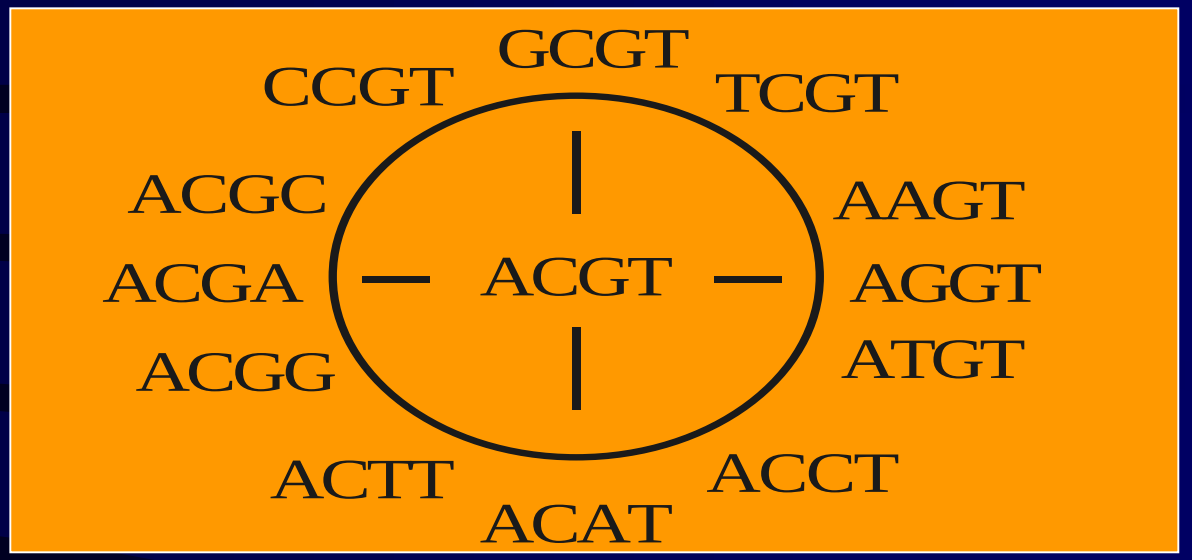
- Generate Beads with Combinatorial Library of Digial Tags:
- Each tag is a string of 8 words chosen from alphabet of 8 words of 4 bases each.
- Synthesize via 8 stages of resin splitting
- Use FACS readout
- Allows differential analysis

Error-Correction Methods of Information Theory



- **Set C of code words.**
  - Encode each N-vector X using a code word C(X) consisting of a N'-vector whose elements range over the same domain.
- **"Maximum Likelihood Error-Correction" procedure:**
  - Maps code word altered by errors back into the nearest error-free code word.
  - *Increases the length of the encoding by* d = N'-N *to gain error-resiliency*.
- **Optimal Codes with independent, random errors:**
  - Restoration error probability r is inverse exponential function 2^{-cN} of N
  - d = N'-N asymptotically approaches 0 as N grows.

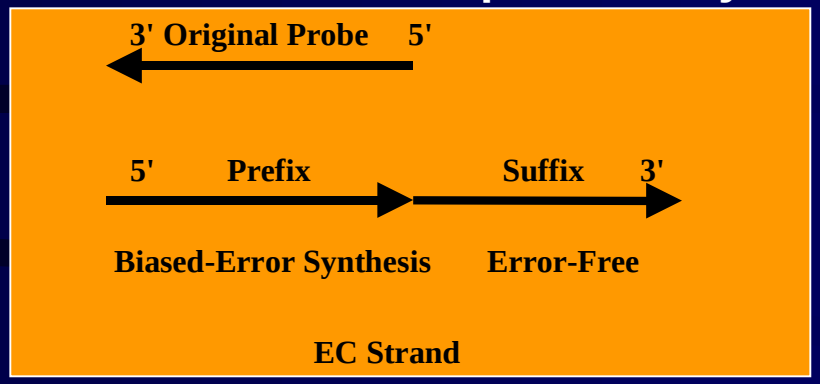
DNA sequences of Hamming distance 1 from ACGT.



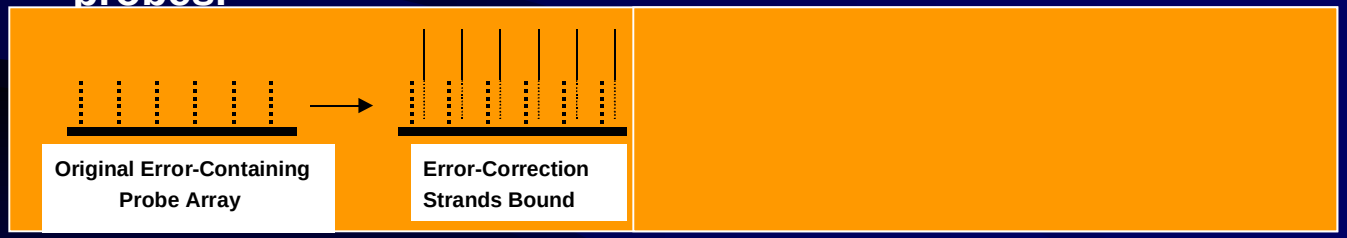
- *2D Projection of a local region in sequence space.*
- *Neighboring sequences are shown for a central tetramer (ACGT) with substitutions in the first position to the north, second to the east, third position south, and fourth west.*

Error-Correction Methods From Computer Science Adapted to Biotechnology

Methods for repairing faulty oligonucleotides contained within surface-bound probe arrays.



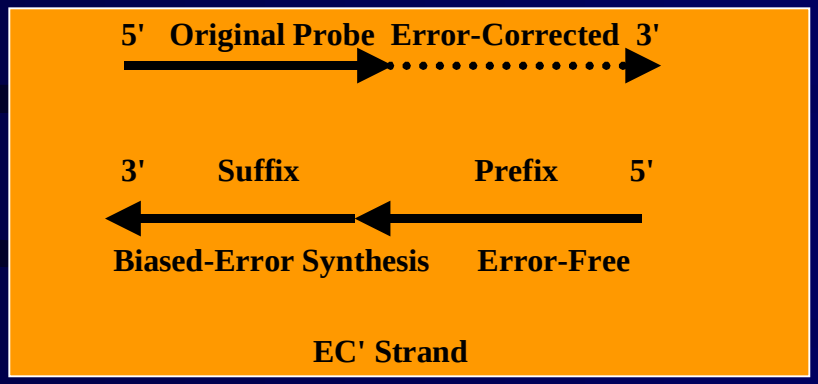
- Use error-correcting codes for design of error-free probes.
- Use "Error-Correction (EC)" DNA strands:
  - Specifically designed to bind both error-containing and error-free probes.



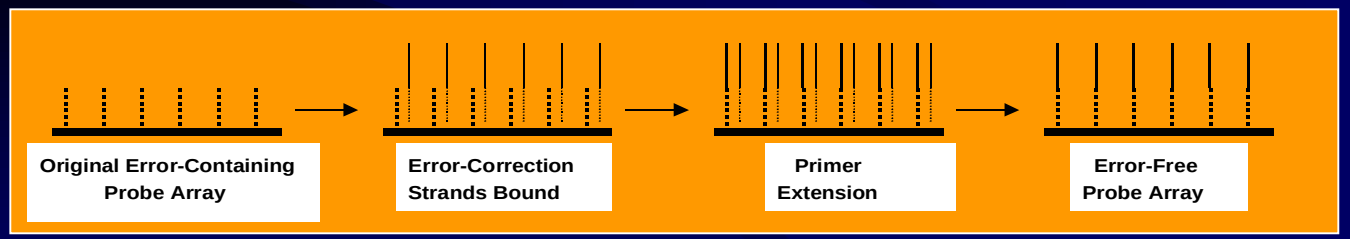
- Error-Correction of Synthesized DNA Strands
  - Resulting in Overhangs.
  - Extra benefit: duplex probes containing single-stranded overhangs less error-prone than simple single-stranded probes.

Error-Correction Methods From Computer Science Adapted to Biotechnology

Error-Corrected (EC) Strand Extension:

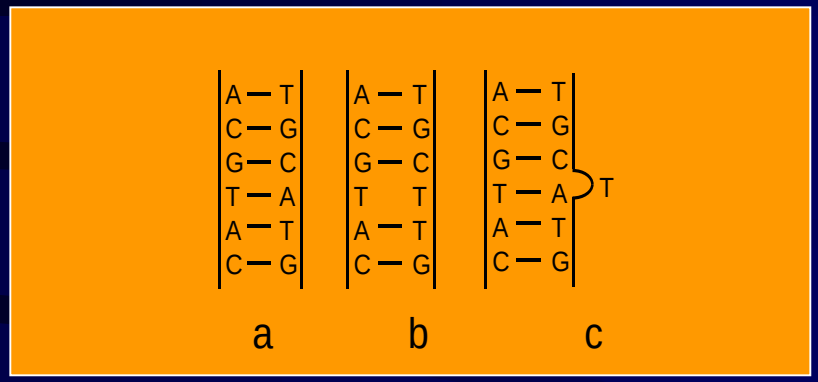


- EC strands can also act as templates for primer extension reactions:
  - append error-free code words onto the 5' ends of all probes on the chip.
  - Strand extending each of the original strands corresponds to its error-free codeword.



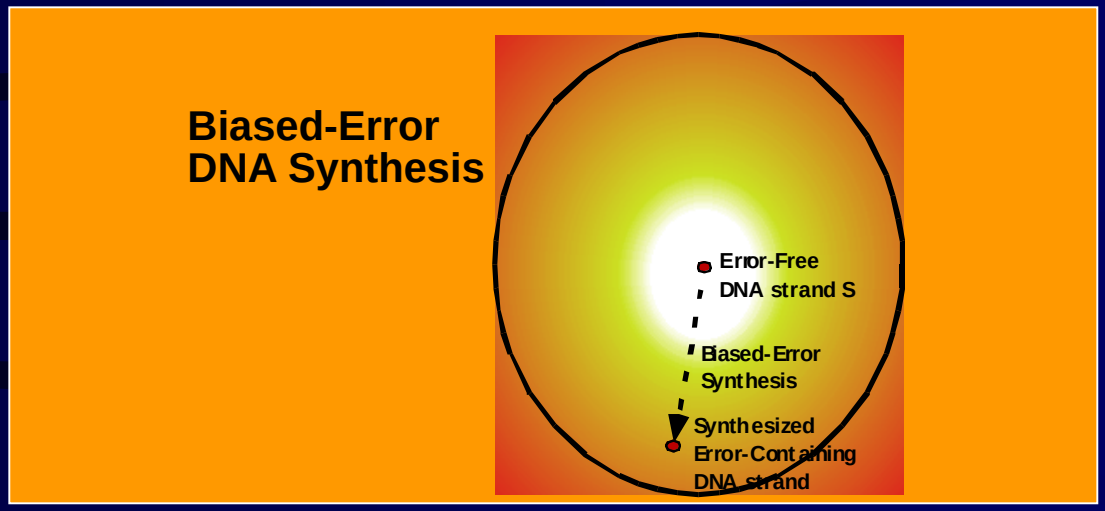
Error Models

Synthesis errors with independent base deletions (causing base bulges) will be first order *approximated by an error model with a uniform, independent probability p of base replacement*



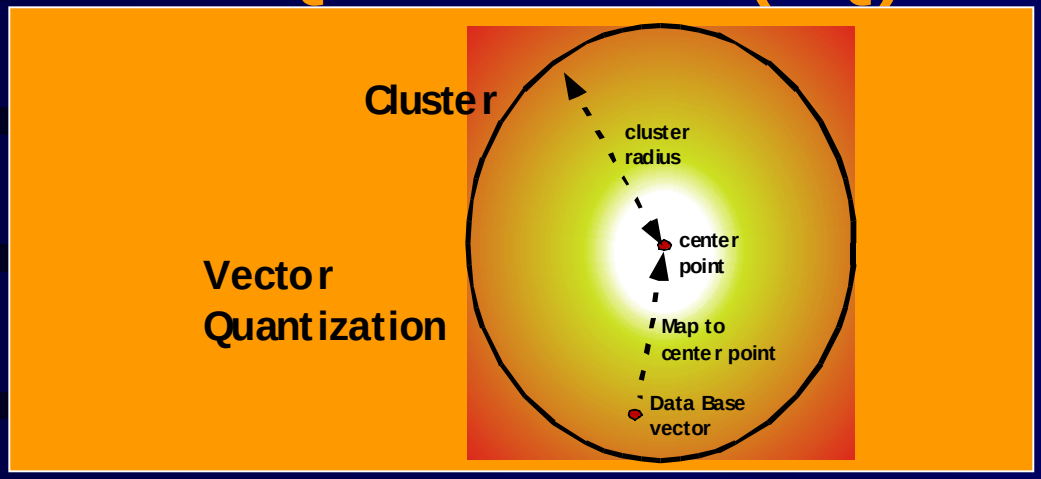
- **Exact and Inexact Hybridization:**
- ***Short stretches of double-stranded DNA are depicted showing:***
  - ***a) exact Watson-Crick(WC) complementary matching;***
  - ***b) a mismatch (T-T) imbedded within a WC match region; and***
  - ***c) a WC match region surrounding a bulged base (T). The bulged base can be described as a deletion from the left-hand strand or an insertion into the right-hand strand.***

Synthesizing EC Strands



- **Direct synthesis and purification:**
  - **small scale only.**
- **Biased-Error Chemical Synthesis:**
  - ***** Recommended (details in paper)**
  - **The relative simplicity of the biased-error chemical synthesis approach makes it the most appealing of methods for generating diverse prefixes for EC strands.**
- **Other Methods for Generating Diverse Prefix for EC Strands:**
  - **Mutagenisis via Polymerase Enzymes.**
  - **DNA Self Assembly.**

Vector Quantization (VQ) Coding



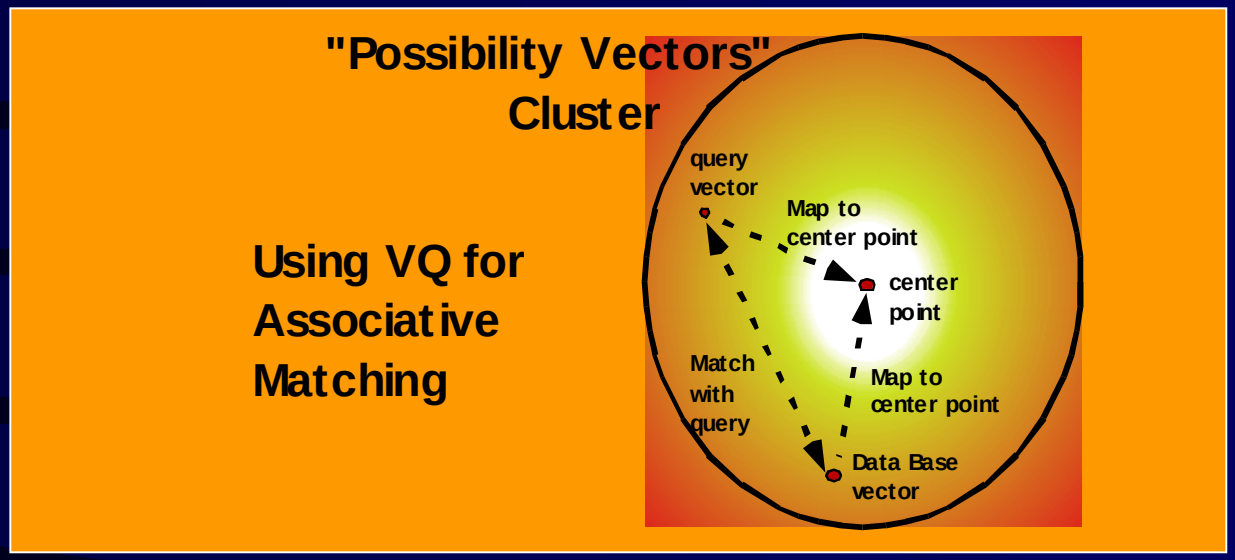
- **Partition vectors of database into clusters of vectors. For each cluster:**
  - **The *center vector* is the average of all the vectors of the cluster.**
  - ***radius* of cluster = maximum distance between any vector of the cluster to center vector.**
  - ***Cluster index* uniquely identifies the cluster.**
- **Well-known algorithms (Jain, Dubes [JD88]) compute clusters:**
  - **Minimizing cluster radius**
  - ***Cluster size* parameter m = average number of vectors in each cluster.**
  - **Number of clusters is a multiple 1/m of original number of vectors of database.**
- **Used in computer science for compressing data (se.g. speech and images) within bounded error.**
  - **Each vector is approximated by the center point of its cluster and coded by the cluster index.**
  - **VQ coding induces errors tuned by choice of parameter m.**
    - **Data-rate/distortion is asymptotically optimal,**
    - **assuming various statistical source models for the data**
    - **(memoryless or finite-state stationary processes [Gray90]).**

Applying VQ Coding Methods

- **To Increase DNA Chip I/O:**
  - Use VQ data clustering techniques to determine the clusters.
  - Only the center points need to be transmitted (at 1/m the cost of transmitting the entire set of the database).
  - Each vector v of the database is represented by a DNA strand encoding:
    - Identification tag for v and
    - Identification tag for center point of cluster containing v.

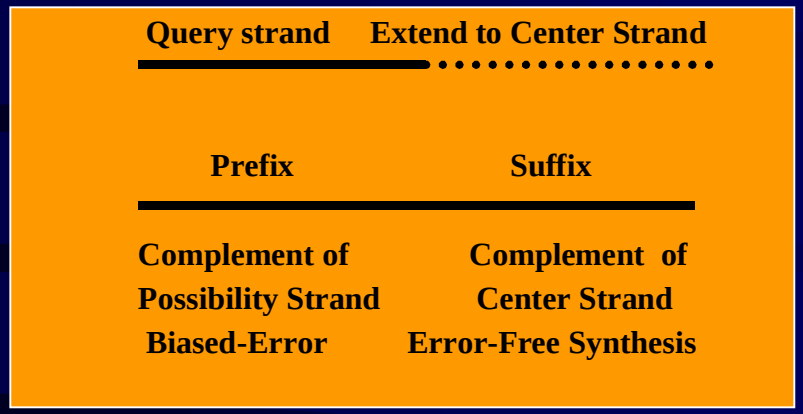
- **Applying VQ Coding Methods To Associative Search: Refining the Associative Search to Exact Matches:**
  - DNA annealing affinity methods work best on complementary sequences.
  - Yet, we need to process an associative match query, even if the query in not an exact match or even partial match with any data in the database.
  - We use VQ-Coding clustering techniques:
    - Reduces associative search problem to finding just exact matches via complementary hybridization.
    - Can be done very effectively by known DNA annealing methods (e.g., PCR).

Reducing Associative Search with Given Match Distance d to the Problem of Exact Match



- For each cluster G of database vectors:
  - The "possibility vectors" of the cluster are those vectors that are within distance d of the center point of G.
  - Query vector v will be included among the "possibility vectors" of those clusters whose centers are of distance at most d from v.
- Vectors in these clusters are at most distance 2d to v, and they include all database vectors that are at most distance d to the query vector, as required.

Associative Search BMC Algorithm using VQ



- **Initialization:**
  - For each cluster G of the database vectors,
    - synthesize a DNA "center stand" of G whose words encode center point of G, along with a unique identification tag.
  - Using a biased-error chemical synthesis, construct from each "center stand", a multiset PS of single stranded DNA, each consisting of a prefix portion that encodes complement of a "possibility vector" of cluster, followed by a suffix portion consisting of complement of "center stand".

- Primer extension is applied; each "query strand" is extended to include "center strand" as its new suffix, forming a "result strand".
- Denature to single stranded DNA, and separate out "result strands" using exact affinity separation with strands complementary to the center strands".
- Output: to conventional media (e.g., by the use of DNA chips) the "center vectors" corresponding to the suffix portions of the "result strands". Determine those cluster vectors of distance at most d from the query vector v.

Associative Query Processing Using VQ

- Input: A query vector v.
- [1] We synthesize a multiplicity of DNA strands, called "query strands" encoding the query vector v.
- [2] Combine these "query strands" with the multiset PS. The hybridization products include doubly stranded DNA complexes. Each of these consist of a "query strand" hybridized with a corresponding prefix of a PS strand, with single stranded overhang consisting of the suffix portion of the PS strand consisting of the complement of the "center strand".
- [3] Primer extension is applied; each "query strand" is extended to include "center strand" as its new suffix, forming a "result strand".
-
- [4] Denature to single stranded DNA, and separate out "result strands" using exact affinity separation with strands complementary to the center strands".
- [5] Output: to conventional media (e.g., by the use of DNA chips) the "center vectors" corresponding to the suffix portions of the "result strands". Determine those cluster vectors of distance at most d from the query vector v.

Extension of Associative Search To Include Boolean Conditionals

- Combine:
  - Our methods for DNA associativesearch with
  - BMC methods for solving the SAT problem
    - (e.g., using surface chemistry techniques).
- Vectors of the database are augmented with "digital tag vectors" consisting of a list of n' Boolean values, encoding binary information about the vector.
- An extended query consist of
  - Query vector to be matched with and
  - Boolean formula to be satisfied.
- The extended query requires finding those database vectors that:
  - closely match the query vector and also
  - whose Boolean variables satisfy the queries Boolean formula.
- Execute the extended query in two stages:
  - First execute the Boolean formula portion of the query as a SAT problem, using biomolecular computing techniques previously developed (e.g., using surface chemistry techniques of Univ. Wisconsin). Strands not encoding SAT solutions are deleted, and all the remaining DNA strands satisfy the Boolean formula.
  - Then execute our associative search procedure on the remaining strands, to find the closest match to the query vector that satisfies the query's Boolean formula.

Conclusion

- The approach of "Computationally Inspired Biotechnologies" may help overcome key biotechnology challenges:
  - increased affinity selectivity for a wider range of molecules [Adelman],
  - decrease errors in chemical synthesis [Reif, LaBean00].
- In the immediate future:
  - Separation between computational and biological technologies should narrow.
- As the miniaturization of biotechnology continues, we can expect DNA chip technology to have also
  - MEMS microflow devices and
  - computational processing capability as well [Gehani and Reif98].