Special course in Computer Science: Molecular Computing

Lecture 5-6: DNA Complementarity and Formal Language Theory

Vladimir Rogojin
Department of IT, Åbo Akademi
http://combio.abo.fi/teaching/special

Fall 2013

Literature recomendations

- Reference on Formal Language Theory and Computability:
  - Arto Salomaa - "Computation and Automata", Cambridge University Press, Mathematics, 2011

DNA encoded data

- DNA/RNA molecules:
  - Data carrier media
  - Alphabet implemented via:
    - Adenine (Uracil, in case of RNA) - A/U
    - Cytosine - C
    - Guanine - G
    - Thymine - T
  - WK complementarity:
    - Antimorphism $\theta(uv)=\theta(v)\theta(u)$:
      - $\theta(A)=T$, $\theta(T)=A$
      - $\theta(C)=G$, $\theta(G)=C$
    - $\theta(w)=\overline{w}$

DNA vs electronic info

- Fundamental difference:
  - In electronic version data is fully controlled
    - Strict addressing
    - Logic and arithmetic operations do not consume operands
    - Normally, unintended operands are not involved into undesired operations
  - In molecular version
    - Complementary DNA can freely hybridize and form unintended structures and implement undesired operations
    - Arguments are consumed by molecular operations
    - Implications:
      - Undesired operations will consume data needed for the designed computation
      - Need to produce enough copies for all the programmed molecular operations

DNA-based information and biooperations

- A DNA molecule usually is represented in millions of identical copies
- The bio-operations usually operate in a massive parallel manner
- Those processes are gorevned by laws of chemistry and thermodynamics
- The output is obeying statistical laws

Encoding of information with DNA

- Designing a set of “good” DNA strands:
  - *Positive* design problem:
    - Design a set of input DNA molecules such that there is a sequence of reactions that produces the correct result.
  - *Negative* design problem:
    - Design a set of input DNA molecules that do not interact in undesirable ways
      - i.e., do not produce incorrect outputs, and
      - do not consume molecules necessary for other “programmed” interactions

DNA complementarity and formal language theory

- Within framework of formal language theory the positive and negative design problems are addressed when creating libraries of oligonucleotides
  - Avoiding intramolecular undesired hybridization
  - Avoiding intermolecular undesired hybridization

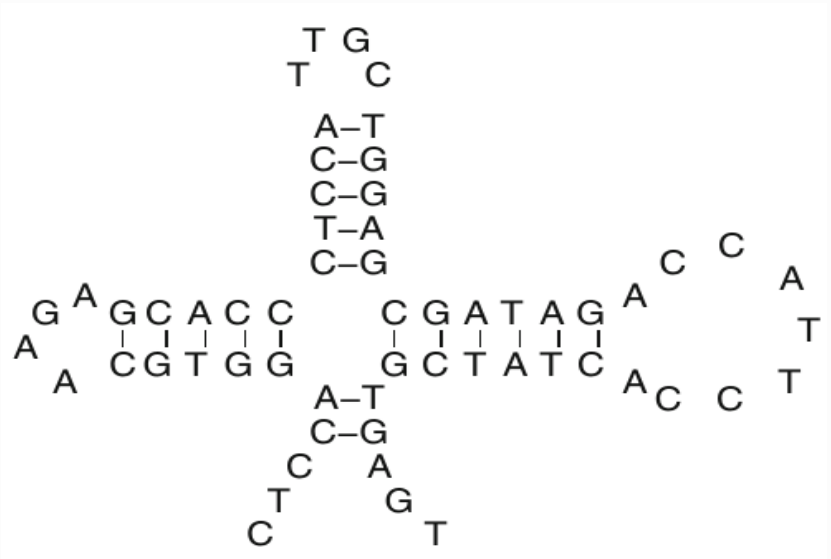
Tube languages

- Tube languages:
  - Formalizes set of molecules in a test tube
  - Tube language *L* is equal to, or a subset of K^+, where *K* is a finite language whose elements are called *codewords*
  - In majority cases *K* consists of words of some fixed length *k*, i.e., *K* – a *code of length k*

Avoiding intramolecular undesired hybridization

- Implemented via hairpin-free words and languages:
  - Θ-k-hairpin-free:
    - u=xvyθ(v)z for some words x, v, y, z implies |v|<k
    - Hairpin structures with stem shorter than k
  - *Θ-k*-scattered-hairpin-free:
    - *u=wy shp(θ,k)*-free if for all words *t*, $t \leq_e w$ and $\theta(t) \leq_e y$ we have $|t| < k$
    - Length of complementary parts in hairpin structure is shorter than k
  - Hairpin frames:
    - Hp-pairs: *(v,θ(v))* in a word u of the form *u=xvyθ(v)z*
    - hp-frame of degree *j* of *u:* $u=x_1v_1y_1\theta(v_1)z_1x_2v_2y_2\theta(v_2)z_2 \ldots x_jv_jy_j\theta(v_j)z_j$

Avoiding intramolecular undesired hybridization



- u=xvyθ(v)z
  - v=GTCAGCGATAG
  - y=ACCATTCCA
- $u=x_1v_1x_2v_2y_2\theta(v_2)y_1\theta(v_1)z$
  - v1=GTCAG, x2=TCA, v2=CGATAG
  - y2=ACCATTCCA, y1=AAC
- $u=x_1v_1y_1\theta(v_1)z_1x_2v_2y_2\theta(v_2)z_2x_3v_3y_3\theta(v_3)z_3x_4v_4y_4\theta(v_4)z_4$
  - *v1=GGTGC, v2=CTCCA, v3=CGATAG*
  - *x1=CTCCA, y1=AAGA, y2=TTGC, y3=ACCATTCCA, z4=TGAGT*

Avoiding intermolecular undesired hybridization

- A number of properties of tube languages prohibiting various types of undesired hybridization between two DNA strands
- Assumption:
  - Hybridization occurs between two perfectly matching complementary subsequences

θ-nonoverlapping

- A language L does not overlap with the language consisting of its complementary words
- In other words: L not intersecting with θ(L)
- Properties:
  - No strand from L will hybridize neither with itself nor with another strand from L

θ-compliant

- For any word *w* from *L*, if w has a complementary match with a subword $\overline{w}$ of *L*, then this subword is a word from *L*
- In other words, *w* does not match to any proper complementary subword in *L*
- Technically, if *w* and *xwy* are words in L, then *x* and *y* are empty words

Θ-prefix/suffix-compliant

- Θ-p-compliant:
  - Any word from L does not match to any proper complementary prefix in L
- Θ-s-compliant:
  - Any word from L does not match to any proper complementary suffix in L

Strictly θ-compliant

- Language L is both θ-compliant and θ-nonoverlapping
- In other words, no word from L has a complementary subword in L(both proper and trivial cases)

θ-Free

- Intersection of L^2 with $\Sigma^+\theta(L)\Sigma^+$ is empty
- In other words, there is no word in L, whose prefix is complementary to a suffix from L and whose suffix is complementary to a prefix from L
- Implication:
  - Two words in L cannot be joined together via the third complementary word

θ-Sticky-free

- For any two words wx and $y\overline{w}$ from L we have x and y empty
- In other words, there are no words in L that could hybridize to each other and form scissors-like free single-stranded tails from both ends

θ-3′/5′-Overhang-free

- θ-3'-overhang-free
  - For any two words wx and $\overline{w}y$ from L we have that x and y are empty words
- θ-5'-overhang-free
  - For any two words xw and $y\overline{w}$ from L we have that x and y are empty words
- Θ-overhang-free
  - Both θ-3'-overhang-free and θ-5'-overhang-free
- Intuition:
  - No two single strands from L can hybridize to form structure with sticky ends from both sides

Subword compliant

- $\theta(k,m_1,m_2)$-subword compliant:
  - There are no words of the form $\Sigma^* u \Sigma^m \theta(u) \Sigma^*$ in L with $|u|=k$ and $m_1 \leq m \leq m_2$
  - Intuitively, L does not contain strands that could form hairpin structures with the stem of length k and hairpin of length between m_1 and m_2

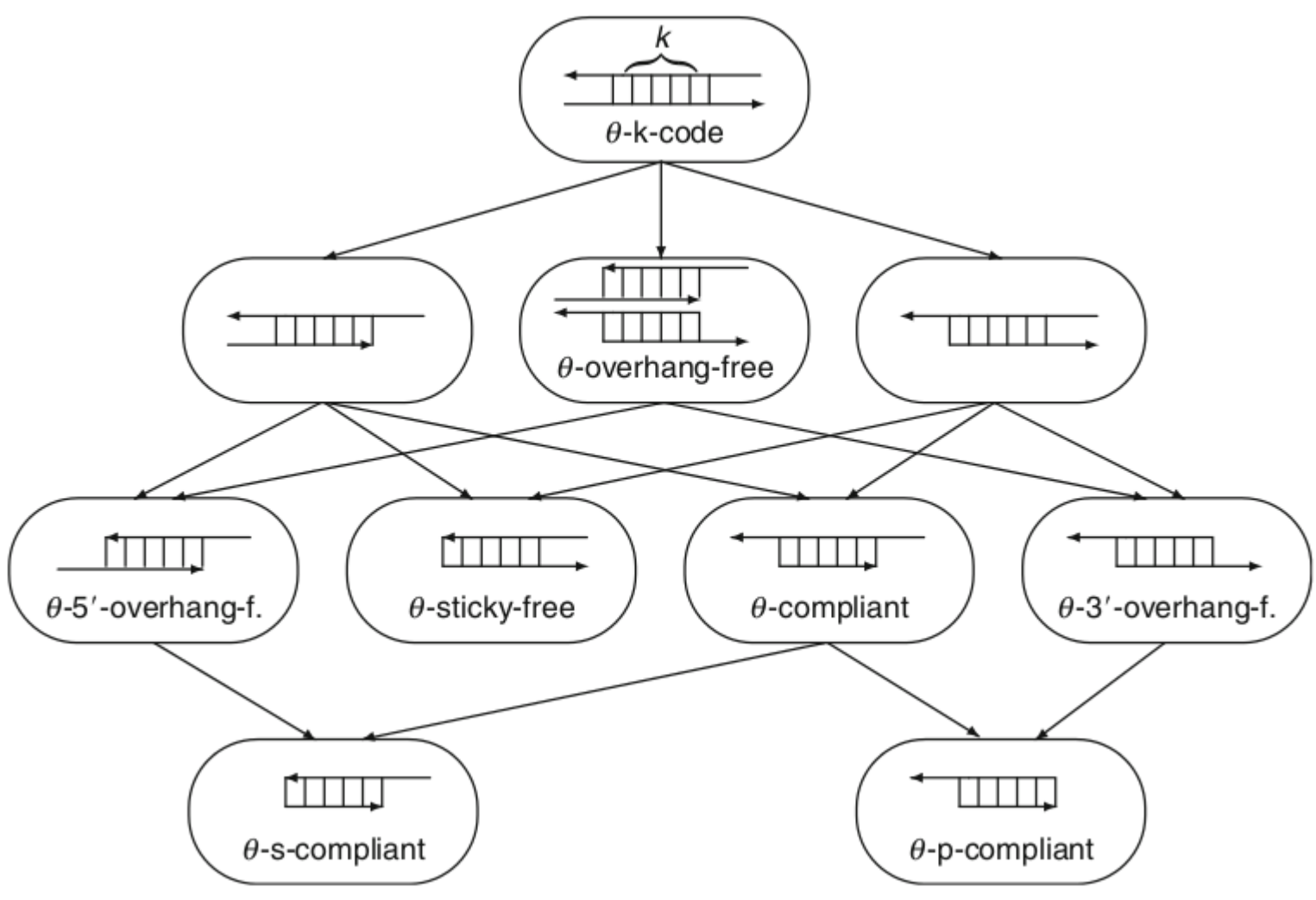
θ-k-Code

- The set of subwords of length k from L does not overlap with the complements of subwords of L of length k
- i.e., sets $Sub_k(L)$ and $Sub_k(\theta(L))$ do not intersect

Solid language

- There is no word in L that is a proper subword from L, and
- There are no two words in L such that a suffix of a words is a prefix of the other word
- i.e., L is a solid language if the following two conditions are satisfied:
  - For all *x, y, z* in Σ^*, where *u, xuy* in L we have that *x* and *y* are empty words
  - For all *x, y* in Σ^* and *u* in Σ^+, where *xu* and *uy* are in *L* we have that *x* and *y* are empty words

Classes of tube languages free of some undesired hybridization



hierarchy of some of the above language properties

Example

- DNA Language $L=\{A^nT^n|n\geq 1\}$ and θ – complementarity automorphism
- L is
  - Not θ-nonoverlapping
  - Not θ-k-code for any positive k
  - Θ-p/s-compliant, since for any w where w, $\overline{w}y/y\overline{w}$ in L it follows that $w=A^nT^n$ and $y=\lambda$
  - Not θ-free, since $A^nT^nA^mT^m$ for n,m>1 is both in L^2 and in $\Delta^+L\Delta^+$
  –

Example (cont.)

- Not θ-sticky-free, as for $w=y=A^n$ and $x=T^n$, and thus we have wx and $y\overline{w}$ in L
- Θ-overhang-free as
  - wx, $\overline{w}y$ in L implies $w=A^nT^m$, $x=T^{n-m}$, $y=T^{m-n}$ and hence $xy=\lambda$, and
  - xw, $y\overline{w}$ in L implies $w=A^nT^m$, $x=A^{n-m}$, $y=A^{m-n}$ and hence $xy=\lambda$
- Not $\theta(k,m_1,m_2)$-subword compliant for any k,m_1,m_2
-

Imperfect DNA-DNA bonds

- So far we have considered perfect binding of complementary DNA strands
- In reality, thermodynamical laws allow for hybridization of “roughly” complementary molecules
- We consider language properties describing and preventing imperfect hybridization

Imperfect DNA-DNA bonds

- Negative design problem is computationally hard if using thermodynamical methods
- Solution: use approximative methods relying on discrete similarity metrics:
  - Hamming distance:
    - the number of positions at which the corresponding symbols are different. Applicable onto two words of same length
  - Levenshtein distance:
    - the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other

Tube languages

- Tube languages:
  - Formalizes set of molecules in a test tube
  - Tube language *L* is equal to, or a subset of K^+, where *K* is a finite language whose elements are called *codewords*
  - In majority cases *K* consists of words of some fixed length *k*, i.e., *K* – a *code of length k*

X[d,k]-property of a K-code

- If u and v are any codewords in K, then $H(u,\theta(v))>d$
- Uniqueness property:
  - $H(K)>d$, where $H(K)$ is the smallest hamming distance between any two words in K

Y[d,k]-property of a tube language

- Let $L=K^{+}$. Here K – is the code of length k and L is a tube language
- If x is a subword of L of length k and v is a codeword in K then $H(x,\theta(v))>d$
-

Z[d,k]-property of a tube language

- If x and y are any subwords of L of length k then $H(x,\theta(y))>d$

Imperfect vs. perfect complementarity

- Property X is generalization of θ-nonoverlaping property:
  - H(u,θ(v))>d vs L∩θ(L)=∉ (u and v are any codewords in K)
- Property Y is generalization of θ-compliant property:
  - H(x,θ(v))>d vs *x*, *xwy* in *L* implies *xy* – empty word (x is a subword of L of length k and v is a codeword in K)
- Property Z generalizes θ-k-Code property:
  - H(x,θ(y))>d vs $Sub_k(L) \cap Sub_k(\theta(L)) = \notin$ (x and y are any subwords of L of length k)

Hybridization of two strands of different lengths, example

Two DNA molecules in which the parts 50 À AAGCGTTCGA À 30 and 50 À TCGGACGTT À 30 bind together although these parts have different lengths.

Hybridization of two strands of different lengths

- One might argue that parts of two DNA molecules could form a stable bond even if they have different lengths
- Based on this observation, the condition for two subwords x and y to bind together should be
  - $|x|, |y| \geq k$ and $Lev(x, \theta(y)) \leq d$,
  - Where Lev(u,v) denotes Levenshtein distance between words u and v
-