

Plasmid-DNAgram : 녹색형광단백질 발현 Plasmid DNA 기반 분자컴퓨팅에 의한 언어 퍼즐 문제 해결*

서울대학교 컴퓨터공학부,¹ 서울대학교 대학원 인지과학 협동과정²
김수동² · 이은석² · 장병탁^{1,2}

Plasmid-DNAgram : Anagram Solving by Molecular Computing Based on GFP-Expressing Plasmid DNA

Su Dong Kim,² Eun-Seok Lee,² Byoung-Tak Zhang^{1,2}

School of Computer Science & Engineering,¹ Seoul National University, Seoul,
Interdisciplinary Program in Cognitive Science,² Seoul National University, Seoul, Korea

요 약

인간 게놈 프로젝트가 완료됨에 따라 생체서열과 언어 사이의 대응 관계가 부각되고 있다. 본고에서는 Lewis Carroll의 언어 유희 사례를 컴퓨터생물학의 측면에서 재조명하고, Carroll이 제시한 문제 중에서 간단한 anagram 문제의 해결을 다루고자 한다. 우선 DNA 컴퓨팅의 방법론을 적용한 DNAgram의 개념을 확장하여 plasmid-DNAgram의 개념을 새롭게 도입하였다. 이 개념을 형광단백질에 대한 FRET (fluorescent resonance energy transfer) 분석기법의 응용 사례인 cameleon 형광단백질에 대한 FRET 분석기법에 적용함으로써, anagram 문제의 어휘론적, 구문론적, 의미론적, 화용론적 측면에 대응하는 바이오분자 컴퓨팅 방법론을 제안하였다.

서 론

인간 게놈 프로젝트(Human Gemome Project, HGP)가 완료됨에 따라,¹⁾ 이 과정에서 얻어진 염기서열을 “生命的書(Book of Life)”라 이름짓고, 이것을 원전(textus receptus)으로 삼아 주해를 달고자 하는 시도를 성경 주해에 빗대어 “생체서열 주해(biosequence exegesis)”라고까지 일컫기에 이르렀다.²⁾

David B. Searls는 생체서열과 언어 사이의 대응 관계에 주목하여, 분자생물학(molecular biology)의 개념 구조를 이루는 ‘서열(sequence)’과 ‘구조(structure)’ 및 ‘기능(function)’에 생리학적 맥락에서 거대분자의 ‘역할(role)’이라는 계층을 더하고, 이에 대응하는 컴퓨터언어학(com-

putational linguistics)의 개념 구조로서 ‘어휘(lexicon)’와 ‘통사(syntax)’ 및 ‘의미(semantics)’ 그리고 ‘화용(pragmatics)’을 제시한 바 있다(Fig. 1).³⁾

본고에서는 언어 유희를 만들어내는데 남달랐던 Lewis Carroll의 다양한 시도를 컴퓨터생물학(computational biology)의 측면에서 재조명하고, Carroll이 제시한 문제 중에서 간단한 anagram 문제 사례에 대하여 DNA 컴퓨팅의 방법론을 적용한 DNAgram의 개념을 확장하여 plasmid-DNAgram의 개념을 도입하며, 이를 형광단백질에 대한 FRET 분석기법의 응용 사례인 cameleon 단백질에 대한 FRET 분석기법에 적용함으로써, anagram 문제의 어휘론적, 구문론적, 의미론적, 화용론적 측면에 대응하는 생물학적 방법론을 제시하고자 한다.

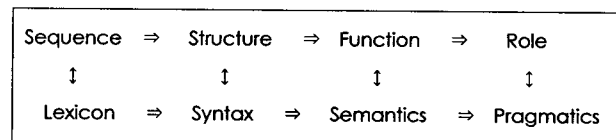


Fig. 1. 분자생물학과 컴퓨터언어학의 대응 개념구조.

*본 연구는 NRL, BrainTech, MEC 과제에 의해 지원되었음. 이 연구를 위해 연구장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에 감사드립니다.

E-mail : sdkim@bi.snu.ac.kr

E-mail : eslee@bi.snu.ac.kr

E-mail : btzhang@bi.snu.ac.kr

Lewis Carroll의 언어 유희 사례

영국의 수학자 Charles Lutwidge Dodgson(1832~1898)은 “이상한 나라의 앨리스(*Alice in Wonderland*)”의 작가로서, Lewis Carroll이라는 필명으로 널리 알려져 있다.^{4,5)} 그는 언어학(linguistics), 기호논리학(symbolic logic), 정수론(number theory), 암호학(cryptography), 확률 이론(probability theory), 기하학(geometry) 등에 심취하여, 기이한 이야기나 비범한 언어 유희, 기발한 수수께끼 등을 문학 작품에 등장시키기에 이르렀다.⁶⁾

Lewis Carroll이 고안한 여러 가지 언어 유희는 오늘날 컴퓨터생물학(computational biology) 방법론의 측면에서 재조명할 가치가 충분하다.⁷⁾ 이 장에서는 Carroll의 언어 유희가 컴퓨터생물학 방법론의 측면에서 잘 구현되고 있음을 몇 가지 사례를 통해 살펴보고자 한다.

1. Doublets

1878년, Carroll은 그가 *Doublet*이라 명명한 퍼즐을 고안해내었다.⁸⁾ 길이가 똑같은 두 단어가 주어지면, 놀이 참가자는 한 번에 오직 한 글자씩만 바꾸되, 중간 단어로서는 모두 올바른 영어 단어만으로 이루어지도록, 처음 제시되었던 두 단어를 서로 이어나가야 한다. 예컨대, ‘ape’와 ‘man’은 다음과 같이 연결될 수 있다(Fig. 2).

Carroll이 제시한 답은 여섯 단계를 거치게 되어 있는데(Fig. 2A), 실은 다섯 단계만을 거치는 경로도 가능하다(Fig. 2B). Doublet은 이렇게 주어진 두 단어에 대하여 최단 경로를 알아내면 이기는 놀이이다. 문자열 사이의 최단 경로 탐색 개념은, 컴퓨터생물학에서 쓰이는 계통발생 재구성 알고리즘(phylogenetic re-construction algorithm)과 서열 비교 분석의 기본이 되는 개념이다.

Carroll의 퍼즐과 생물학적 서열 분석 사이의 명백한 차이점은, 생물학의 경우, “올바른” 중간 서열이나 조상 서열이 알려져 있지 않으므로, 사전처럼 참고할 자료가 없다는 점이다. 또 다른 차이점이라면, Doublet 퍼즐은 치환만

| |
|---------------------------------|
| (A) APE→ARE→ERE→ERR→EAR→MAR→MAN |
| (B) APE→APT→OPT→OAT→MAT→MAN |

Fig. 2. Doublet의 첫째 사례 : ‘ape’와 ‘man’.

| |
|--------------------------|
| IRON→ICON→COIN→CORN→CORD |
| →LORD→LOAD→LEAD |

Fig. 3. Doublet의 둘째 사례 : ‘iron’과 ‘lead’.

을 허용하며, 생물학적 서열 비교분석 알고리즘에서는 상대적으로 드물기 때문에 벌점(penalty)과 함께 반영되는, 삽입(insertion)과 缺失(deletion)을 허용하지 않는다.

1892년, Carroll은 새로운 규칙을 고안하여 한 단어 안에서 문자들이 자유로이 교환될 수 있게 하였다.⁴⁾ 따라서, ‘iron’에서 ‘lead’로 이르는 방법은 다음과 같다(Fig. 3).

이 경우, 완화된 규칙이 적용되지 않으면 Doublet은 불가능해 보인다. Donald Knuth는 다섯 글자로 이루어진 5,757개의 영어 단어를 대상으로 연구하여, 671개 단어는 어느 것과는 연결되어 있지 않으며, 103쌍은 오로지 쌍쌍끼리만 연결되어 있음을 보였다.⁴⁾ 이러한 자료구조에서 Doublet(shortest-path graph algorithm)으로 귀결되며, 전체 그래프(overall graph)에서 서브 그래프(subgraph)를 찾는 것과 연관된다. 이것은 서열 유사성이나 단백질-단백질 상호작용, 공통된 발현 양상 등에 근거하여 유전자들의 네트워크를 나타내는 것과 유사하다.

2. Segmentation

Carroll은 Syzygies에서 공백(gap)을 허용하는 것을 고려하지 않았는데, 그 까닭은 의심할 나위 없이 이러한 공백은 문제를 일으킬 소지가 다분할 뿐만 아니라, 줄줄이 이어지는 단어 내부에 존재할 경우 적어도 미관상 좋지 않기 때문이다. 그럼에도 불구하고 그는, 다른 환경에서는 심지어 공백을 이용하도록 만들기까지 했다. 앞서 완화된 Doublet으로 제시된, “iron”에서 “lead”로 변환시키는 예를 보건대(Fig. 3), 그가 추가적인 변환기작을 고려하지 않았음은 명확해 보인다. 그는 문자의 자유로운 치환, 즉 “철자 바꾸기(anagrammatization)”를 택했다. 오늘날, anagram은 대중적인 인기를 끌어 모으고 있는데, 문자의 재배열뿐만 아니라 자유로이 공백을 삽입 또는 재배치하여 새로운 문구를 만들어내기도 한다. 예컨대, “dormitory”로부터 “dirty room”이라는 anagram이 만들어진다. Carroll 또한, 미발표분을 포함하여 수많은 anagram을 만들어 내었는데, 여기에는 당대의 저명인사의 이름에 기초한 것도 상당수 포함된다.⁹⁾ 예를 들자면, 1868년 영국의 총리로 취임한 William Ewart Gladstone의 이름에 기초한 anagram은 다음과 같다. “Wild agitator! Means well.”

Carroll은 또한 다음과 같이 공백의 문제를 교묘히 다룬 유명한 단어 퍼즐을 고안해내었다.⁴⁾

“Rearrange the letters of ‘nor do we’ to make one word.”

알려진 바에 따르면, 이 문제에 대한 답은 없다. 즉, 주어진 문자로 구성된 단일 단어는 없지만, “one word”라는 문구 자체가 “nor do we”의 anagram이 된다. 따라서 문제에 이미 답이 포함되어 있지만, 얼핏 보아서의 답은 눈에 띄지 않는데, 전형적인 Carroll 방식의 역설에 의하여 숨겨져 있다.

이렇듯 Carroll이 다룬 문제들은 종종 몇몇 가지 종류의 비균일한 서열을 구분하는 컴퓨터생물학의 문제로 귀결된다. 이러한 사례들은 공백 할당과 경계 식별 문제의 이중성에 대한 문제와 함께 조합의 특색을 드러낸다. Carroll이 제기한 것과 같은 종류의 모호성의 문제들은 DNA의 경우에서도 발견되는데, 유전자가 엑손(exon)과 인트론(intron)으로 나뉘어 선택 접합(alternative splicing)과 같은 현상에 내재되어 있을 뿐만 아니라, 길게 연속된 유전체 서열에서 유전자의 경계를 정확히 결정하는 문제에서도 적용된다.¹⁰⁾ 단백질의 경우에는, 1차 서열로부터 α -나선구조나 β -병풍구조와 같은 2차 구조를 예측 하는 프로그램을 만들거나,¹¹⁾ 도메인(domain)의 경계를 확정하는 프로그램을 만들 경우, 이와 비슷한 문제에 제작자들이 직면하게 된다.¹²⁾ 생물학에서의 이러한 구분 문제의 일반성은 효율적인 해법을 추구하는 알고리즘 제작자들에게 명확히 인식되고 있다.¹³⁾

3. Nonsense

Carroll은 Edward Lear와 함께, *Nonsense verse*로 알려진 문학 형태의 주요 선구자였는데, 그 전형적인 모습은 “*The Mad Gardner’s Song*”에서 발췌한 일부분을 통해 살펴볼 수 있다.⁷⁾

He thought he saw a Rattlesnake
That questioned him in Greek :
He looked again, and found it was
The Middle of Next Week.
“The one thing I regret,” he said,
‘Is that it cannot speak!’

He thought he saw an Argument
That proved he was the Pope :
He looked again, and found it was
A Bar of Mottled Soap.
‘A fact so dread,’ he faintly said,
‘Extinguishes all hope!’

부조리에서 익살을 발견하는 노력을 기울이는 한편으로, Carroll은 또한 통사론(syntax)과 의미론(semantics) 사이의 분리 현상을 탐구하였다. 그 뒤 100년도 지나지 않아 Noam Chomsky는, “Colorless green ideas sleep furiously”라는 문장을 통사적으로는 올바르지만 의미론적으로는 무의미한 발화의 예시로 이용하여 심각한 과학적 문제를 제기하였고,¹⁴⁾ 최근 두뇌 조영술(brain imaging)을 이용한 연구를 통해 언어의 이러한 측면의 분리 현상이 보고되었다.¹⁵⁾ Carroll의 Nonsense verse는 잘 구성되었을 뿐만 아니라, 우아하기까지 하지만, 실제 세계의 어떤 현상과도 관련시킬 수 없을 만큼 일관성을 결여하고 있다. 언어학에서와 마찬가지로, 생물학에서도 통사론 및 의미론의 구분이 이용되고 있는데, 예컨대 실제 올바르게 기능할 수 있는 단백질을 암호화하고 있는 유전자(gene)에 대해, 類似유전자(pseudogene)는 통사론 및 의미론 측면에서 올바르지 않다고 할 수 있다. 생물학자들은 심지어 유전자 중간에 종결 코돈(stopcodon)이 끼어들어 불가능성 폴리펩티드(polypeptide)를 만들게 되는 경우에 대해 *nonsense mutation*이라는 용어를 사용한다.

Plasmid-DNAgram

본고에서는, 2.에서 제시한 바 있는 anagram 문제를 바이오분자 컴퓨팅으로 해결하는 문제를 고찰하고자 한다. 이은석¹⁶⁾은 anagram 문제를 DNA 컴퓨팅으로 해결하는 개념을 제시하고, 이를 “DNA-gram”이라는 프로그램을 통해 구현한 바 있다. 본고에서는, 이 개념을 더욱 확장, anagram 문제를 해결하기 위하여 plasmid DNA 조작을 통해 FRET (fluorescence resonance energy transfer) 현상을 바이오분자에 적용¹⁷⁾하여 해결하는 방안을 제시하고자 하며, 이를 한글 및 한국어 정보처리에 적용해보고자 한다.

1. Plasmid-DNAgram : 어휘론

Carroll의 anagram 문제를 DNA 컴퓨팅으로 구현하려면 알파벳 철자 각각을 DNA 염기서열에 대응시키는 절차가 필요하다. 본고에서는 실험적 편의를 위하여, 알파벳 철자를 각기 DNA restriction site에 대응시키는 경우를 다루고자 한다. 또한, 보다 진전된 실험에 적합하도록 알파벳 정보에 대응하는 DNA 염기서열이 plasmid vector의 MCS (multi-cloning site)에 삽입되는 경우를 상정해 두고자 한다.

알파벳 철자를 DNA restriction site에 대응시키면, restriction map 작성을 통해 문자열의 구성과 순서를 확인할 수 있다. 마찬가지로, 한글 자모를 DNA restriction

site에 대응시키면, 한글자모열의 구성과 순서를 restriction map 작성을 통해 확인해 낼 수 있다.

간단한 예로서, 알파벳 철자 “D”를 제한효소의 일종인 *Hind*III 인식부위에 대응시키는 경우를 생각해 보자. Fig. 4와 같은 가상적인 상황에서, 제한효소의 또 다른 일종인 *Bam*H I 절단부위를 양쪽에 sticky end로 가지는 1.3kb 크기의 DNA 조각은, *Bam*H I site를 포함하는 plasmid DNA의 MCS에 삽입시킬 수 있다.¹⁸⁾ Insert DNA 조각에는 *Hind*III 인식부위뿐만 아니라 여러 가지 restriction site 들을 포함시킬 수 있으므로, insert DNA 조각에 대하여 restriction map을 구성하는 과정은 알파벳 철자의 어순을 확인하는 과정에 대응한다.

확장된 예로서, 알파벳 철자 “E”를 제한효소 *Eco*R I 인식부위(5'-GAATTC-3')에, 알파벳 철자 “N”을 제한효소 *Kpn* I 인식부위(5'-GGTACC-3')에, 그리고 알파벳 철자 “O”는 제한효소 *Xho* I 인식부위(5'-CTCGAG-3')에 각각 대응시키는 경우를 생각해 보자. Insert DNA 조각에 대한 restriction map 구성결과, 5'-말단으로부터 제한효소 인식부위의 순서가 “*Eco*R I - *Xho* I - *Kpn* I”인 경우는 “EON”, “*Xho* I - *Kpn* I - *Eco*R I”인 경우는 “ONE”, 그리고 “*Kpn* I - *Eco*R I - *Xho* I”인 경우는 “NEO”에 해

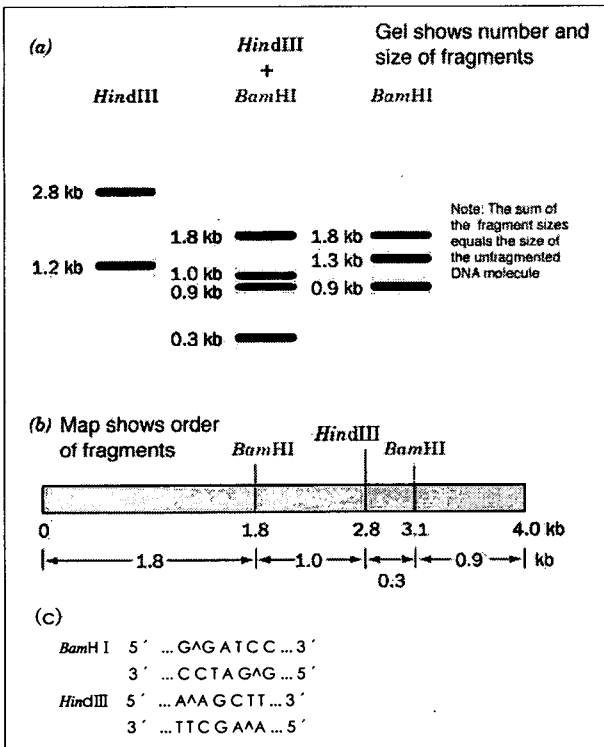


Fig. 4. Restriction mapping. A : DNA 제한효소 *Bam*H I 또는 *Hind*III에 의한 DNA 전기영동패턴. B : A에서 주어진 정보를 통해 작성된 restriction map. C : DNA 제한효소 *Bam*H I 또는 *Hind*III의 DNA 인식 및 절단부위 염기서열.

당한다.

녹색형광단백질(GFP, green fluorescent protein)을 발현하는 plasmid인 GFP-expression vector를 이용하면, insert DNA 조각에 포함된 제한효소 인식부위 의 순서에 대응하는 알파벳 철자의 배열이 올바른 단어를 나타내는지 여부에 관한 메타 정보를 표현할 수 있다.¹⁹⁾ Fig. 5에 나타난 p35S-GFP와 같은 GFP-expression vector는, *E. coli*에 주입되면 GFP를 생산하여 녹색 형광을 발현한다. 따라서, 올바른 단어가 주어질 경우, 이에 상응하는 재조합 DNA를 이용하여 형광단백질을 발현시킬 수 있다.

2. 바이오분자에서의 FRET 활용

FRET(Fluorescence resonance energy transfer)은 분자간 거리를 측정하기 위한 “분광학적 척도(spectroscopic ruler)” 역할을 해왔다.²⁰⁾ 분자생물학에서 GFP를 이용하게 되면서 FRET을 실험기법으로 적용할 수 있게 되었다.¹⁷⁾

FRET의 원리는 고준위 에너지 전달체 역할을 하는 발광체로부터 저준위 에너지를 지닌 수용체 분자로 직접적으로 에너지 전달이 이루어져, 수용체 분자를 들뜨게 하여 전달체의 형광을 감쇄시키는 데 있다.

전달체(donor)와 수용체(acceptor) 분자들은 반드시 전달체의 방출 파장(emission wavelength)과 수용체의 들뜸 파장(excitation wavelength)이 겹쳐지도록 짝지어져야 한다. 하지만, 전달체와 수용체의 파장이 과도하게 겹쳐지지 않도록 하여야 FRET 신호 측정이 용이해진다. FRET을 측정하기 위해서는, 전달체가 들뜰 수 있는 범위 안에서 시료를 적절한 파장의 빛으로 쬐이면, 수용체로부터 방출되는 빛의 파장을 측정할 수 있다.

에너지 전달의 효율(E)은 전달체와 수용체 발광소자의 거리(r)에 크게 의존하며, Forster equation으로 기술된다

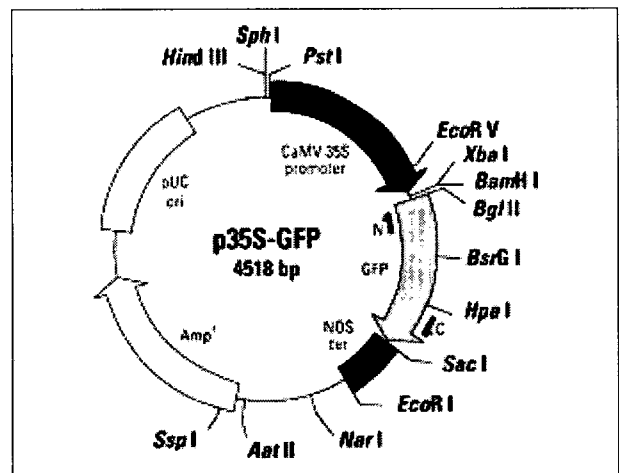


Fig. 5. p35S-GFP의 restriction map.

[수식 1].²¹⁾ Förster 반경(R_0)은 각 형광 분자의 상대적 위치에 따라 분광학적 성질로부터 계산된 최대값의 50% 효율만큼 에너지 전이가 이루어지는 거리이다.

$$E = \frac{R_0^6}{R_0^6 + r^6}$$

[수식 1] Förster equation

예컨대, 수용체와 전달체 사이의 거리가 $R=R_0$ 로부터 $R=2R_0$ 로 두 배 늘어날 경우, 에너지 전달의 효율(E)은 50%에서 1.5%로 줄어든다. 따라서, FRET은 분자 내부 거리의 작은 변화에도 매우 민감한 측정값을 제공한다. 대부분의 경우, FRET은 100 Å 이상의 거리에서는 측정할 수 없으므로, 생물학적으로 단백질의 상호작용을 연구하는데 있어서 분자간 근접도를 알려주는 훌륭한 척도가 된다.¹⁷⁾

해파리의 일종인 *Aequorea victoria*로부터 GFP가 얻어진 이래, 그의 돌연변이 형태인 BFP(blue fluorescent protein), CFP(cyan fluorescent protein), YFP(yellow fluorescent protein) 등이 개발되었으며, 이들 중에서 FRET에 가장 적합한 두 쌍은 BFP-GFP 쌍과 CFP-YFP 쌍이다(Fig. 6).^{17,22)} 최근 *Discosoma* 屬(genus)의 산호류로부터 RFP(red fluorescent protein)가 발견되어 GFP-RFP 쌍 또한 FRET에 활용할 수 있게 되었다.²¹⁾

3. Plasmid-DNAgram에서 형광단백질 발현의 의미 : 구문론 및 의미론

1.에서 구축한 insert DNA를 상기해 보자. 우리는 이미 영어 단어 “one”을 나타내는 DNA 서열을 고려한 바 있다.

마찬가지 방법으로 영어 단어 “word”에 대응하는 DNA 서열을 고안할 수 있다.

추가로, 알파벳 철자 “D”를 제한효소 *Dra* I 인식부위(5'-TTTAAA-3')에, 알파벳 철자 “R”을 제한효소 *EcoR* V 인식부위(5'-GATATC-3')에, 그리고 알파벳 철자 “W”는 제한효소 *BsiW* I 인식부위(5'-CGTACG-3')에 각각 대응시키면, “word”에 대응하는 DNA 서열 “*BsiW* I - *EcoR* V - *Dra* I”을 얻을 수 있다.

“one word”의 two word anagram 중 하나인 “new door”의 경우를 생각해 보자. 영어 단어 “new”에는 “*Kpn* I - *EcoR* I - *BsiW* I”, 영어 단어 “door”에는 “*Dra* I - *Xho* I - *Xho* I - *EcoR* V”가 각각 대응한다.

세 글자 단어인 “one”과 “new”에는 각각 CFP에 덧붙인 단백질 X_{one} , X_{new} 를 대응시키고, 네 글자 단어인 “word”와 “door”에는 각각 YFP에 덧붙인 단백질 Y_{word} , Y_{door} 를 대응시키면, CFP는 구문론적 자질 중에서 형용사에 대한 메타 정보를 나타내며, YFP는 구문론적 자질 중에서 명사에 대한 메타 정보를 나타낸다.

또한, 의미론적으로 올바른 구문을 구성하는 단백질 X_{one} 과 단백질 Y_{word} 의 결합 또는, 단백질 X_{new} 과 단백질 Y_{door} 의 결합에 의해서만 관찰되는 FRET 현상은 의미론적 자질에 대응한다(Fig. 6A, B).²¹⁾

4. Plasmid-DNAgram에서 cameleon 형광단백질 발현조절의 의미 : 화용론

3.에서 우리는 “one word”의 two word anagram 중 하나인 “new door”의 경우를 살펴보았다. 이 절에서는 “one word”의 two word anagram 중 또 다른 하나인 “new

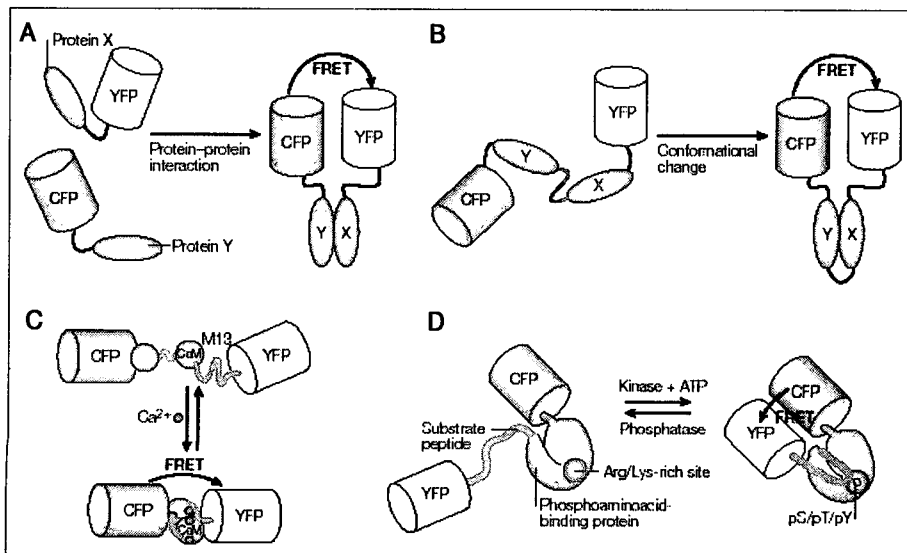


Fig. 6. 두 가지 형광단백질을 이용한 바이오분자-FRET의 개념도.

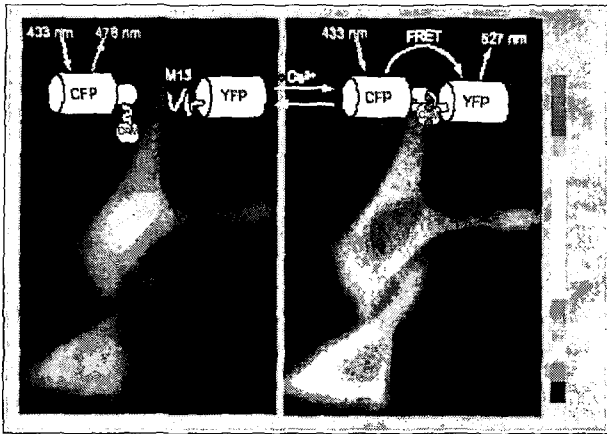


Fig. 7. Ca^{2+} 이온농도 변화에 따른 CFP/YFP FRET.

odor”의 경우를 생각해 보자. 영어 단어 “odor”에는 DNA 서열 “Xho I - Dra I - Xho I - EcoRV”가 대응한다.

Two word anagram “new door”와 “new odor”는 모두 구문론적으로나 의미론적으로 결함이 없는 구문이다. 그러나 話用論的 수준(pragmatic level)에서 보면, “new door”와 “new odor”는 활용빈도의 측면에서 차이가 있다고 할 수 있다.

Chimeric protein의 일종인 cameleon 단백질은 “BFP/CFP-CaM(calmodulin)-glycylglycine linker-M13(the CaM CaM-binding domain of myosin light chain kinase)-GFP/YFP”의 차례로 연결된 구조로 구성되어 있으므로 FRET에 이용할 수 있다.²³⁾

Ca^{2+} 이온이 CaM에 결합하면 CaM과 M13을 분자 내부에서 서로 결합하게 한다. 그 결과, cameleon 분자 내부를 보다 촘촘하게 만드는 구조적 변화가 일어나 cameleon 분자의 양쪽 말단에 존재하는 두 가지 형광단백질로 이루어진 단백질쌍 사이의 FRET 효율을 높여준다(Fig. 6C).^{21,23,24)} 세포 내부의 Ca^{2+} 이온농도가 높은 경우 강한 FRET이 관찰되지만, 세포 내부의 Ca^{2+} 이온농도가 낮은 경우 FRET은 미미하게 나타날 따름이다(Fig. 7).^{23,24)} 세포 내부의 Ca^{2+} 이온 농도 대신 PKA(protein kinase A) 또는 PKC의 단백질 활성 등 다양하게 분포하는 세포 내부의 변수를 이용하여 FRET을 관찰하는 방법이 알려진 바 있다(Fig. 6D).^{21,25)}

세포 내부의 Ca^{2+} 이온농도와 화용론적 수준에서의 활용빈도를 대응시킬 경우, 높은 Ca^{2+} 이온 농도는 상대적으로 활용빈도가 높은 “new door”라는 구문을, 낮은 Ca^{2+} 이온 농도는 상대적으로 활용빈도가 낮은 “new odor”라는 구문을 나타낸다고 할 수 있으며, cameleon FRET의 측정값을 구문에 대한 활용빈도의 값에 대응시킬 수 있다.

결론 : Plasmid-DNAgram을 이용한 한글 및 한국어 처리

지금까지 우리는 영문자를 사용한 anagram 문제를 다루었다. 한글을 이용한 anagram 문제도 마찬가지로 해결할 수 있다. 한글에 적용되는 예제로서는 박목월 시인의 시, ‘나그네’에서, ‘구름에 달 가듯이’라는 부분을 발췌하여 사용해보자.²⁶⁾

‘구름’이라는 단어는 자음 ‘ㄱ’, ‘ㄹ’, ‘ㅁ’ 및 모음 ‘ㅜ’, ‘ㅡ’의 조합으로 이루어져 있다. 이들 한글 자모는 각기 다른 DNA restriction site에 대응시킬 수 있으며, 이를 재조합 하면, ‘그물’이라는 단어도 만들 수 있다. 같은 방법으로, ‘구름에 달 가듯이’라는 시구 전체를 한글 자모로 풀어헤친 후 재조합 해보면, ‘달에 구름 가듯이’, ‘구름이듯 달 가에’, ‘달이듯 구름 가에’, ‘그 강 다 물들이 세’, ‘그 시(詩) 강에 물들다’, ‘그 물에 강(江) 시들다’, ‘그물(網)에 강시(疆屍) 들다’ 등등 다양한 어구를 만들 수 있다. 구절마다 나타나는 단어들을 각기 재조합 plasmid DNA에 부호화하여 단어마다 서로 다른 형광단백질에 대응시키고, 발현되는 형광 패턴 분석을 통하여 주어진 어구가 나타내는 다양한 화용론적 의미를 대응시킬 수 있다.

본고에서는 이상과 같이 DNA 컴퓨팅의 방법론을 적용한 DNAgram의 개념을 확장하여 plasmid-DNAgram의 개념을 새롭게 도입하였다. 이 개념을 cameleon 단백질에 대한 FRET 분석기법에 적용함으로써, anagram 문제의 어휘론적, 구문론적, 의미론적, 화용론적 측면에 대응하는 바이오파자 컴퓨팅 언어처리 방법론을 제안하였다.

향후 연구로 지금까지 제안된 바이오파자 컴퓨팅 방법론을 실험을 통해 실제로 구현할 예정이다.

REFERENCES

- 1) Francis S. Collins, Michael Morgan, Aristides Patrinos(2003) : *The Human Genome Project : Lessons from Large-Scale Biology*. Science 300 : 286-290
- 2) Mark S. Boguski(1999) : *Biosequence Exegesis*. Science 286 : 453-455
- 3) David B. Searls(2001) : *Reading the Book of Life*. Bioinformatics 17 (7) : 579-580
- 4) Martin Gardner(1996) : *The Universe in a Handkerchief : Lewis Carroll's Mathematical Recreations, Games, Puzzles, and Word Plays*. Copernicus/Springer-Verlag, New York
- 5) Lewis Carroll(1936) : *The Complete Works of Lewis Carroll*. Modern Library, New York. Reprinted in paperback by Penguin, London, 1999
- 6) Morton N. Cohen(1995) : *Lewis Carroll : A Biography*. A.A. Knopf, New York
- 7) David B. Searls(2001) : *From Jabberwocky to Genome : Lewis Carroll and Computational Biology*. J Comput Biol 8 (3) : 339-348
- 8) Edward Wakeling(1992) : *Lewis Carroll's Games and Puzzles*,

- Dover, New York
- 9) Edward Wakeling(1995) : *Rediscovered Lewis Carroll Puzzles*, Dove, New York
 - 10) Roderic Guigo, Pankaj Agarwal, Josep F. Abril, Moises Burset, James W. Fickett(2000) : *An assessment of gene prediction accuracy in large DNA sequences*. *Genome Res* 10 (10) : 1631-1642
 - 11) Scott C. Schmidler, Jun S. Liu, Douglas L. Brutlag(2000) : *Bayesian segmentation of protein secondary structure*. *J Comp Biol* 7 (1-2) : 233-248
 - 12) Sarah J. Wheelan, Aron Marchler-Bauer, Stephen H. Bryant(2000) : *Domain size distribution can predict domain boundaries*. *Bioinformatics* 16 (7) : 613-618
 - 13) Gary D. Stormo, David D. Haussler(1994) : *Optimally parsing a sequence into different classes based on multiple types of evidence*. *ISMB* 2 : 369-375
 - 14) Noam Chomsky(1957) : *Syntactic Structures*. Mouton, the Hague
 - 15) Mirella Dapretto, Susan Y. Bookheimer(1999) : *Form and Content : Dissociating Syntax and Semantics in Sentence Comprehension*. *Neuron* 24:427-432
 - 16) 이은석(2003) : *DNAgram : Anagram 문제 해결에 관한 분자 컴퓨팅 시뮬레이션 연구*. 서울대학교 대학원 석사학위 논문
 - 17) Richard M. Siegel, Francis Ka-Ming Chan, David A. Zacharias, Ruth Swofford, Kevin L. Holmes, Roger Y. Tsien, Michael J. Lenardo(2000) : *Measurement of Molecular Interactions in Living Cells by Fluorescence Resonance Energy Transfer Between Variants of the Green Fluorescent Protein*. *Sci. STKE* 2000 (38), p11. at http://www.stke.org/cgi/content/full/OC_sigtrans;12000/38/p11
 - 18) Donald Voet, Judith G. Voet(1990) : *Biochemistry*. John Wiley & Sons, Inc., New York
 - 19) Martin Chalpie, Y. Tu, G. Euskirchen, WW. Ward, DC. Prasher(1994) : *Green Fluorescent Protein as a marker for gene expression*. *Science* 263 : 802-805
 - 20) Stryer L(1978) : *FluoRescence Energy Transfer as a spectroscopic ruler*. *Annu Rev Biochem* 47 : 819-846
 - 21) Jin Zhang, Robert E. Campbell, Alice Y. Ting, Roger Y. Tsien(2002) : *Creating New Fluorescent Probes for Cell Biology*. *Nat Rev Mol Cell Biol* 3 (12) : 906-918
 - 22) Kevin Truong, Mitsuhiro Ikura(2001) : *The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo*. *Curr Opin Struct Biol* 11 (5) : 573-578
 - 23) Atsushi Miyawaki, Oliver Griesbeck, Roger Heim, Roger Tsien(1999) : *Dynamic and Quantitative Ca²⁺ Measurements Using Improved Cameleons*. *Proc Natl Acad Sci USA* 96 : 2135-2140
 - 24) Paul R. Selvin(2000) : *The Renaissance of Fluorescent Resonance Energy Transfer*. *Nat Struct Biol* 7 (9) : 730-734
 - 25) Jin Zhang, Yuliang Ma, Susan S. Taylor, Roger Y. Tsien(2001) : *Genetically encoded reporters of Protein Kinase A activity reveal impact of substrate tethering*. *Proc Natl Acad Sci USA* 98(26) : 14997-15002
 - 26) 박목월(1975) : *구름에 달 가듯이*. 삼중당, 서울