ライフサイエンス分野におけるテキストマイニング技術適用の動向

武藤　克弥

富山県立大学　電子・情報工学科

July 9, 2021

はじめに

背景

テキストマイニングはテキストデータを様々な観点から分析し，構造化することにより有益な情報を得る技術であり，自然言語処理，情報検索・抽出，データマイニングなどを組み合わせている．近年ではテキストマイニングをバイオ・医療といったライフサイエンス分野に用い，文献・データベースから遺伝子やたんぱく質の関係性を見い出す研究が活発となっている．

目的

- ある遺伝子配列を文献や医療データベース中に現れる遺伝子配列と比較し，類似度や構造関係を抽出する技法を紹介する
- 代謝ネットワークを可視化し，そのグラフ構造を見直す手法を紹介する

(1) 遺伝子マッチング

(1) 遺伝子マッチングの概要

ある遺伝子配列について医療データベースや文献と照合し，類似の配列を抽出したり，複数の配列どうしの因果関係・構造関係を見出す研究は古くから行われており，再現率や適合度などを用いた評価がなされてきた.

研究の目的

- データベースからの抽出に時間がかかることがしばしば問題視されていた
- BLAST という遺伝子 DB 管理・類似性検索ツールを使えば，従来の 50 倍の速さで実用的な準最適解を求めることができる

  →生物学分野のある論文内の遺伝子・タンパク質配列を BLAST がどれだけ抽出できるか検証する

(1) 遺伝子マッチング2

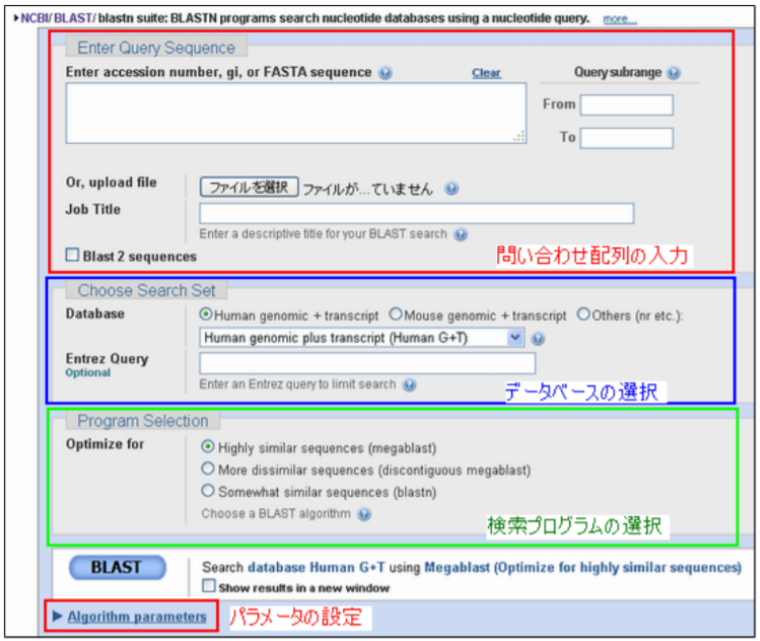


図 1: BLAST 検索欄

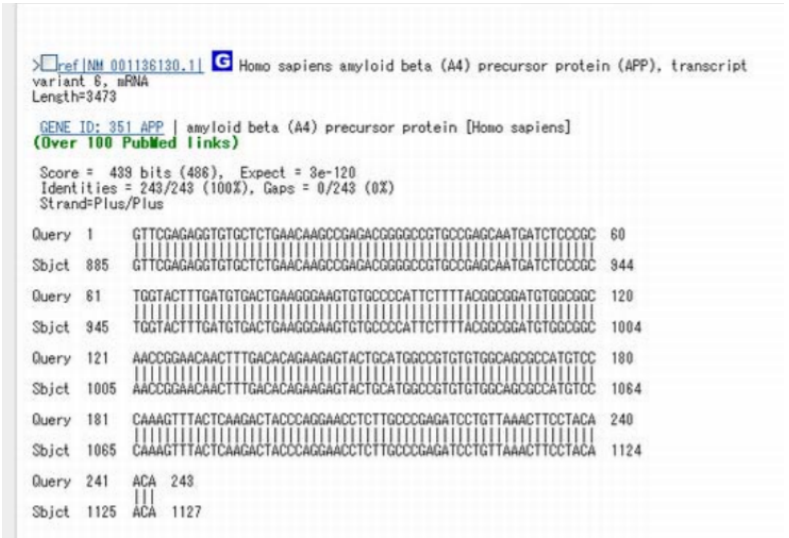


図 2: 類似度比較

遺伝子配列の抽出方法

A	AAAC	E	AACG	I	AAGT	M	ACAC	Q	ACCG	U	ACGT	Y	AGAG
B	AAAG	F	AACT	J	AATC	N	ACAG	R	ACCT	V	ACTC	Z	AGAT
C	AAAT	G	AAGC	K	AATG	O	ACAT	S	ACGC	W	ACTG		
D	AACC	H	AAGG	L	AATT	P	ACCC	T	ACGG	X	ACTT		

0	AGCC	4	AGGG	8	AGTT	/	ATCC	,	ATCC	?	ATCC	-	ATCC
1	AGCG	5	AGGT	9	ATAT	\	ATCC	;	ATCC	"	ATCC		
2	AGCT	6	AGTC	]	ATCC	(	ATCC	:	ATCC	.	ATCC		
3	AGGC	7	AGTG	[	ATCC	)	ATCC	!	ATCC	space	ATCC		

図 3: 変換表

```
For instance, ErbB, Ras, and Raf all lie on the ERK MAP
kinase (MAPK) pathway
```

```
AACTACATACCTATCCAAGTACAGACGCACGGAAACACAGAAATAACGATCCATCCAAC
GACCTAAAGAAAGATCCATCCACCTAAACACGCATCCATCCAAACACAGAACCATCCAC
CTAAACAACTATCCAAACAATTAATTATCCAATTAAGTAACGATCCACATACAGATCCA
CGGAAGGAACGATCCAACGACCTAATGATCCACACAAACACCCATCCAATGAAGTACAG
AAACACGCAACGATCCATCCACACAAACACCCAATGATCCATCCATCCACCCAAACACG
GAAGGACTGAAACAGAG
```

図 4: 論文内文章の配列化

検証と結果

Database[a]	*E*-value		Wordsize[b,c]		Alignments		Penalty	
	Optimized	Default	Optimized	Default	Optimized	Default	Optimized	Default
3	10	10	12 (3)	11 (<3)	2000	250	−6	−3
4–5	10	10	16 (4)	11 (<3)	2000	250	−6	−3
6–10	10	10	20 (5)	11 (<3)	250	250	−3	−3
11–20	10	10	40 (10)	11 (<3)	250	250	−3	−3
>20	10	10	80 (20)	11 (<3)	250	250	−3	−3

図 5: パラメータ設定

Parameter settings	Names marked by evaluators and included in database(n=753)			Names marked by evaluators and not included in database (n=409)			Names not marked by evaluators
	True positive		False negative	True positive		False negative	False positive
	Full match	Partial match	No match	Full match	Partial match	No match	
Optimized	712 (94.6%)	23 (3.1%)	18 (2.4%)	18 (4.4%)	163 (39.9%)	228 (55.7%)	362
Default	424 (56.3%)	69 (9.2%)	260 (34.5%)	18 (4.4%)	104 (25.4%)	287 (70.2%)	214

図 6: 検証結果

(2) 遺伝子構造の可視化

研究の目的

- タンパク質や RNA、低分子といった生化学物質どうしの制御・伝達関係，因果関係を可視化する
- FCM(Fuzzy Cognitive Map) を用いて相互の関係性を明確にする

FCM とは

- ある概念どうしの因果関係をモデル化するもの
- 従来の遺伝子制御ネットワークではフィードバックの部分をモデル化できなかった
  →それに対応するようにした

FCM の概要

FCM の概要

- あるノードどうしにおいて，正の因果関係を $+1$, 負を -1, 因果関係なしを 0 とし，集合 $\{-1, 0, 1\}$ のエッジ重みで表す
- 入力に因果関係の変化が生じたときにエッジ行列が更新される学習法を用いる ($C_n(t)$ を更新)
- k：上流ノード，i：下流ノード，$S(y)$：シグモイド関数

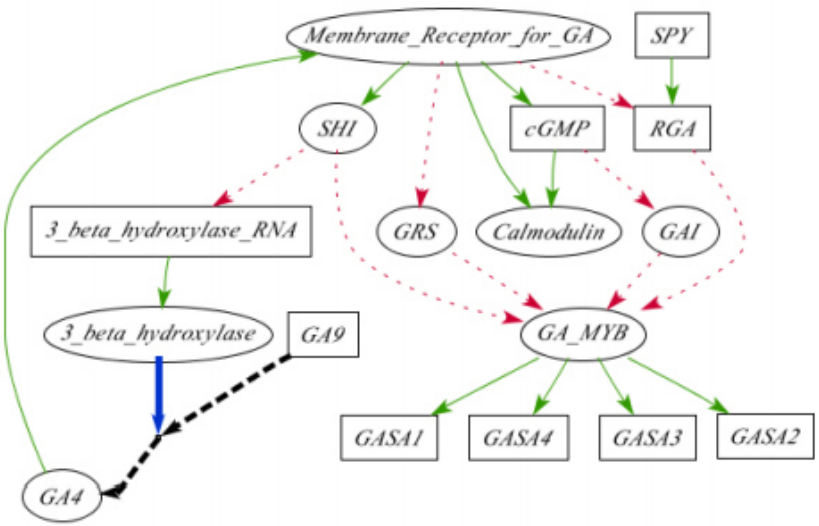
$$C_i(t_{n+1}) = S\left[\sum e_{ki}(t_n) C_k(t_n)\right]$$

$$S_j(y_j) = \frac{1}{1 + e^{-c(y_j - T_j)}}$$

代謝ネットワークの可視化

代謝ネットワークの可視化

システムの検証としてシロイヌナズナにおける植物ホルモンであるジベレリンの代謝とシグナル伝達をモデルに当てはめた



代謝ネットワークの可視化２

代謝ネットワークの可視化２

- 制御リンク
  →上ノードから下ノードへの重みの調整 (影響度に応じて変化)
- 変換リンク
- 触媒リンク
- 強制関数