バイオインフォマティクス
（第2回）

慶應義塾大学生命情報学科
榊原康文

ヒトゲノム解読完了

日本など宣言

ブレア英首相は14日、英国議会で「イラクのほぼ全土が英米軍の支配下に入った」と表明。中央軍前線司令部のブルックス准将は同日の記者会見で、「軍事作戦は転換点に入った」と語った。

ヒトのすべての遺伝情報であるヒトゲノム解読計画で、日米英など6カ国首脳は14日、「解読完了」を宣言した。00年に90%を読みとった「概要版」が発表されたが、科学的には今回が本当の節目。遺伝子は約3万2千個とわかった。

（3面に関係記事）

人の遺伝情報は人体の設計図ともいえ、細胞の核にあるDNAに刻まれている。DNAは4種類の塩基（文字）が鎖状につながってできている。

今回、28億6千万文字のうち技術的に読みとれない約1%を除く99%を解読。精度は99・99%以上で、データは公開される。日本からは理化学研究所や慶応大学などが参加し約1億8500万文字を読みとった。全体の約6%で、米59%、英31%につぐ貢献度だった。病気にかかわる遺伝情報の研究など、生命科学研究の基盤ができたことになる。

ヒトゲノム　ヒトの体は約60兆個の細胞からなり、その一つひとつの核にあるDNA（デオキシリボ核酸）に同じ遺伝情報が記録されている。ゲノムとはこの遺伝情報全体のこと。

大岡　信

グランドにしゃがざる十七歳の老

（朝日新聞
2003年4月15日）

ヒトゲノムの塩基配列：
28億6千万塩基
99%を解読
精度は99．99%以上

年)4月13日　金曜日　朝日新聞

アカゲザル全遺伝情報 解読

感染症研究への応用期待

米欧中などの国際チーム

【ワシントン＝上田俊英】病気の治療研究などに広く利用されているアカゲザルの全遺伝情報（ゲノム）を、米欧や中国などの国際チームが解読した。霊長類のゲノム解読はヒト、チンパンジーに次ぎ3種目。進化の仕組みを探る上で貴重なデータとなる。また、アカゲザルはエイズなどの感染症研究に欠かせない動物で、遺伝子レベルでの研究応用が期待される。13日付の米科学誌サイエンスに論文が載る。

アカゲザルはニホンザルと同じマカク属で、インドから中国にかけてのアジアに広く生息する。アカゲザルとヒトの祖先は、約2500万年前に分かれた。

解読チームはインド系の1匹のメスから取ったゲノムを調べた。遺伝情報がのる染色体は21対あり、ヒトの23対、チンパンジーの24対と若干異なるが、遺伝情報の違いはヒトと約2・5%、チンパンジーとも約2・5%だった。ヒトとチンパンジーの遺伝情報の違いは約1・2%とされる。

体毛の生え方、免疫反応などヒトとの進化の違いに関係しそうな約200の遺伝子も特定。解読チームリーダーで米ベイラー医科大ヒトゲノム解読センター長のリチャード・ギブス博士は「アカゲザルの全遺伝情報は、何がヒトと他のサルとを分けたのかを解明する大きな手がかりになる」と話す。

また、アカゲザルはいまから16万年前ごろ、インド系と中国系とに大きく分かれた。両者を比較すると、例えばサル版のエイズウイルスに感染した際、インド系の方が症状の進行が速いという。こうした免疫機能の違いが、どのような遺伝情報の違いと関係しているのかを探る研究にも役立ちそうだ。

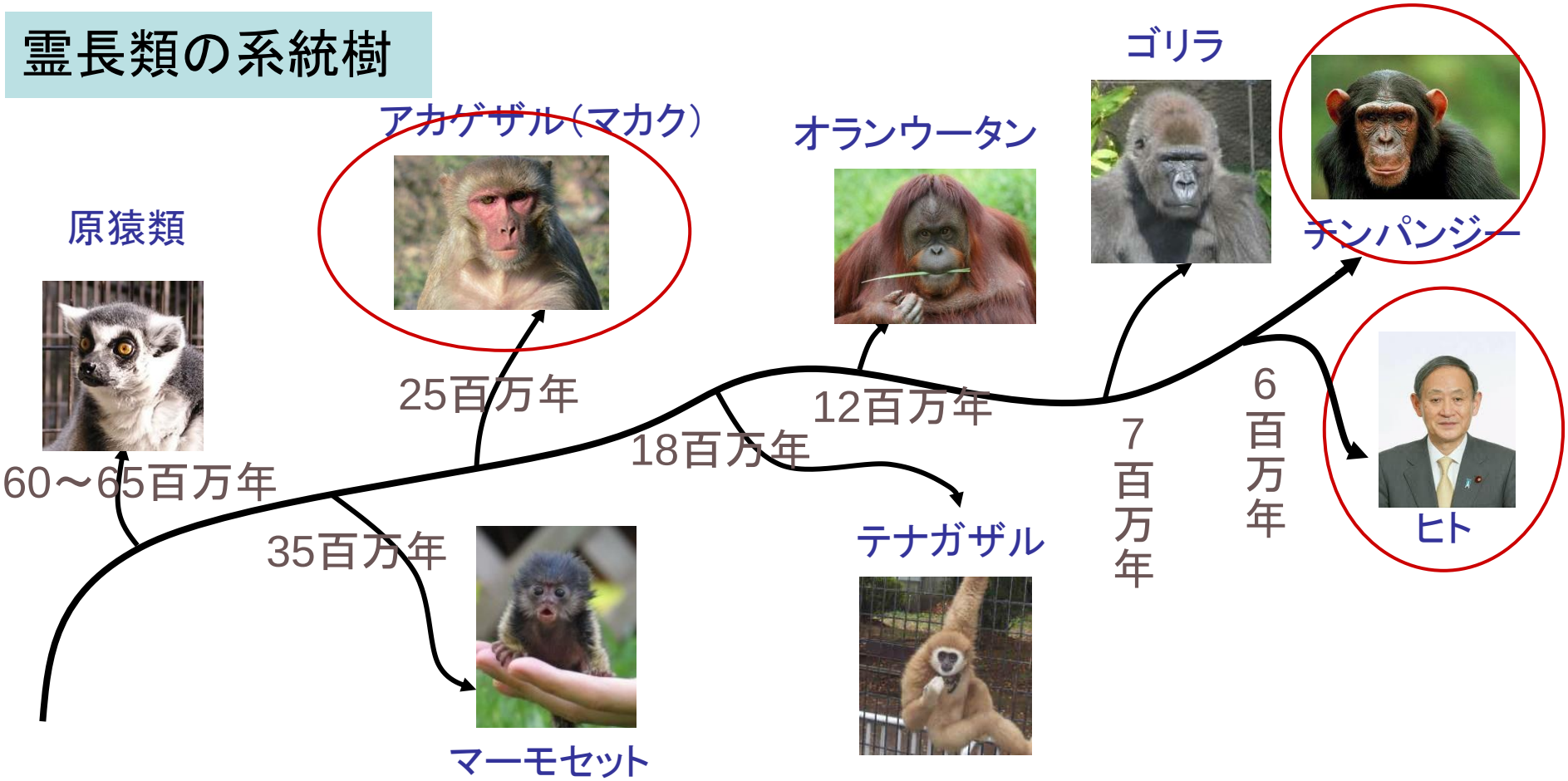
（朝日新聞
2007年4月13日）

アカゲザルのゲノム配列：
- 染色体21対
  （ヒト23対，チンプ24対）
- 遺伝子領域の違い
  ヒトやチンパンジーと
  約2.5%
  （ヒトとチンプの間1.2%）
- などなど

↓
比較ゲノム解析

霊長類ゲノムの比較

霊長類の系統樹



- ヒト（human）
- チンパンジー（chimp）
- アカゲザル（rhesus）

Human 22+X+Y
Chimp 23+X+Y
Rhesus 20+X

マーモセットゲノムの解読

Jayakumar *et al. BMC Genomics* 2020, **21**(Suppl 3):243
https://doi.org/10.1186/s12864-020-6657-2

BMC Genomics

RESEARCH Open Access

An improved de novo genome assembly of the common marmoset genome yields improved contiguity and increased mapping rates of sequence data

Vasanthan Jayakumar[1], Hiromi Ishii[1], Misato Seki[1], Wakako Kumita[2], Takashi Inoue[2], Sumitaka Hase[1], Kengo Sato[1], Hideyuki Okano[3,4], Erika Sasaki[2] and Yasubumi Sakakibara[1*]

アサガオの全ゲノム解読

nature COMMUNICATIONS

ARTICLE

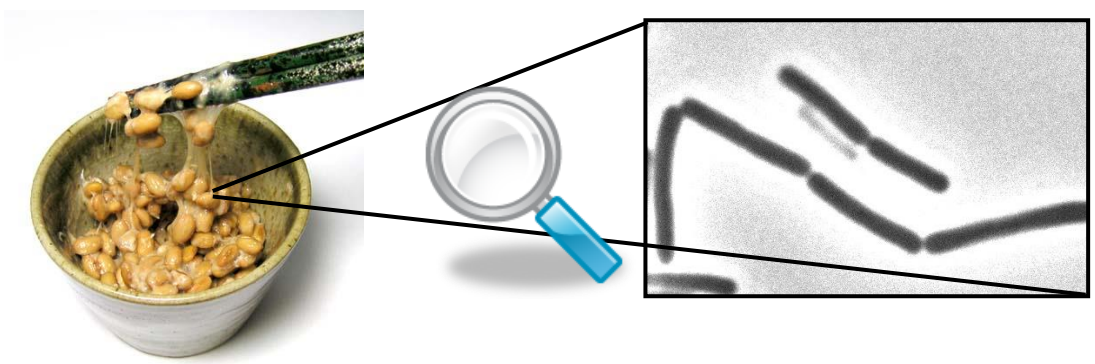
Received 30 Jan 2016 | Accepted 21 Sep 2016 | Published 8 Nov 2016

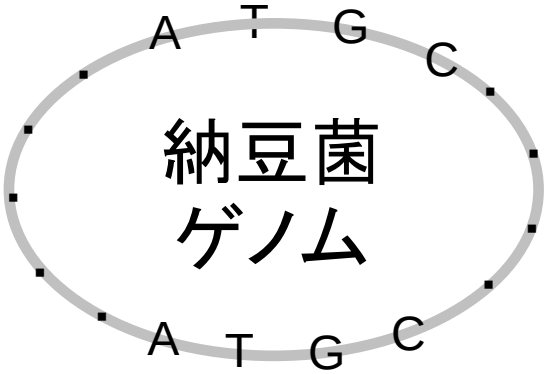
DOI: 10.1038/ncomms13295 OPEN

Genome sequence and analysis of the Japanese morning glory *Ipomoea nil*

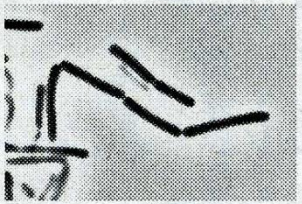
Atsushi Hoshino[1,2,*], Vasanthan Jayakumar[3,*], Eiji Nitasaka[4], Atsushi Toyoda[5], Hideki Noguchi[5], Takehiko Itoh[6], Tadasu Shin-I[5], Yohei Minakuchi[5], Yuki Koda[3], Atsushi J. Nagano[7,8], Masaki Yasugi[7,†], Mie N. Honjo[7], Hiroshi Kudoh[7], Motoaki Seki[9,10], Asako Kamiya[9], Toshiyuki Shiraki[11], Piero Carninci[12], Erika Asamizu[13,†], Hiroyo Nishide[1], Sachiko Tanaka[1], Kyeung-Il Park[1,14], Yasumasa Morita[1,†], Kohei Yokoyama[4], Ikuo Uchiyama[1,2], Yoshikazu Tanaka[15], Satoshi Tabata[13], Kazuo Shinozaki[9], Yoshihide Hayashizaki[16], Yuji Kohara[5], Yutaka Suzuki[17], Sumio Sugano[18], Asao Fujiyama[5,19], Shigeru Iida[1,2,†] & Yasubumi Sakakibara[3]

納豆菌（*Bacillus subtilis* natto ）のゲノム





納豆菌ゲノム解読



納豆を作る納豆菌＝写真、慶応大提供＝のゲノム（全遺伝情報）を、慶応大の榊原康文教授らの研究チームが初めて解読した。納豆のネバネバの秘密は、塩基配列のわずかな違いにあった。

納豆菌は枯草菌の一種でネバネバ成分を作るが、ネバネバを作らない枯草菌もいる。研究チームは、納豆菌のネバネバ作りに関する主な遺伝子二つと、ネバネバを作らない枯草菌の同じ場所にあたる遺伝子を比べた。すると遺伝子がそれぞれ１塩基しか違っていなかった。

納豆菌のゲノムサイズは約409万塩基対。遺伝子数は推定約４千個だった。

（朝日新聞　朝刊科学面
２０１０年５月１４日）

Nishito *et al. BMC Genomics* 2010, **11**:243
http://www.biomedcentral.com/1471-2164/11/243

BMC Genomics

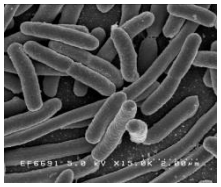
RESEARCH ARTICLE Open Access

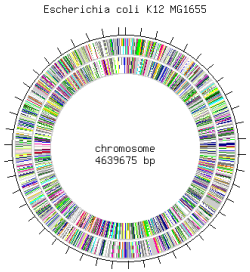
Whole genome assembly of a natto production strain Bacillus subtilis natto from very short read data

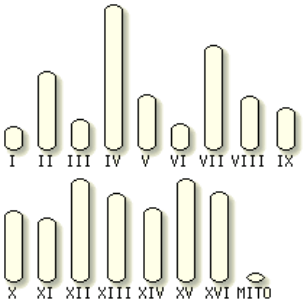
Yukari Nishito[1], Yasunori Osana[2], Tsuyoshi Hachiya[1], Kris Popendorf[1], Atsushi Toyoda[3], Asao Fujiyama[4], Mitsuhiro Itaya[5] and Yasubumi Sakakibara*[1]

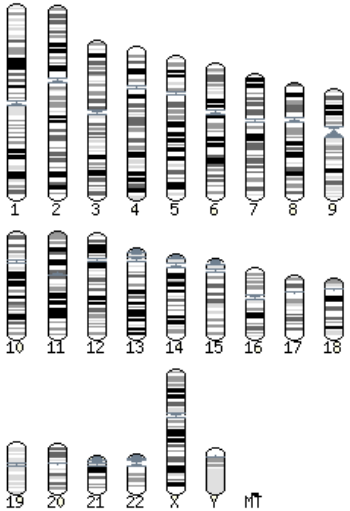
(Nishito et al., *BMC Genomics*, 2010)

ゲノムの大きさ

大腸菌	出芽酵母	ショウジョウバエ	ヒト
			
500万塩基対	1,200万塩基対	1.3億塩基対（1.8億塩基対）	30億塩基対
	2倍	**30倍**	**600倍**







Ensembl Genome Browser
http://www.ensembl.org/index.html

Golden path length
（reference assembly length）

ゲノムの配列決定

① 物理地図（DNAマーカーの位置）の作成

② ショットガン法

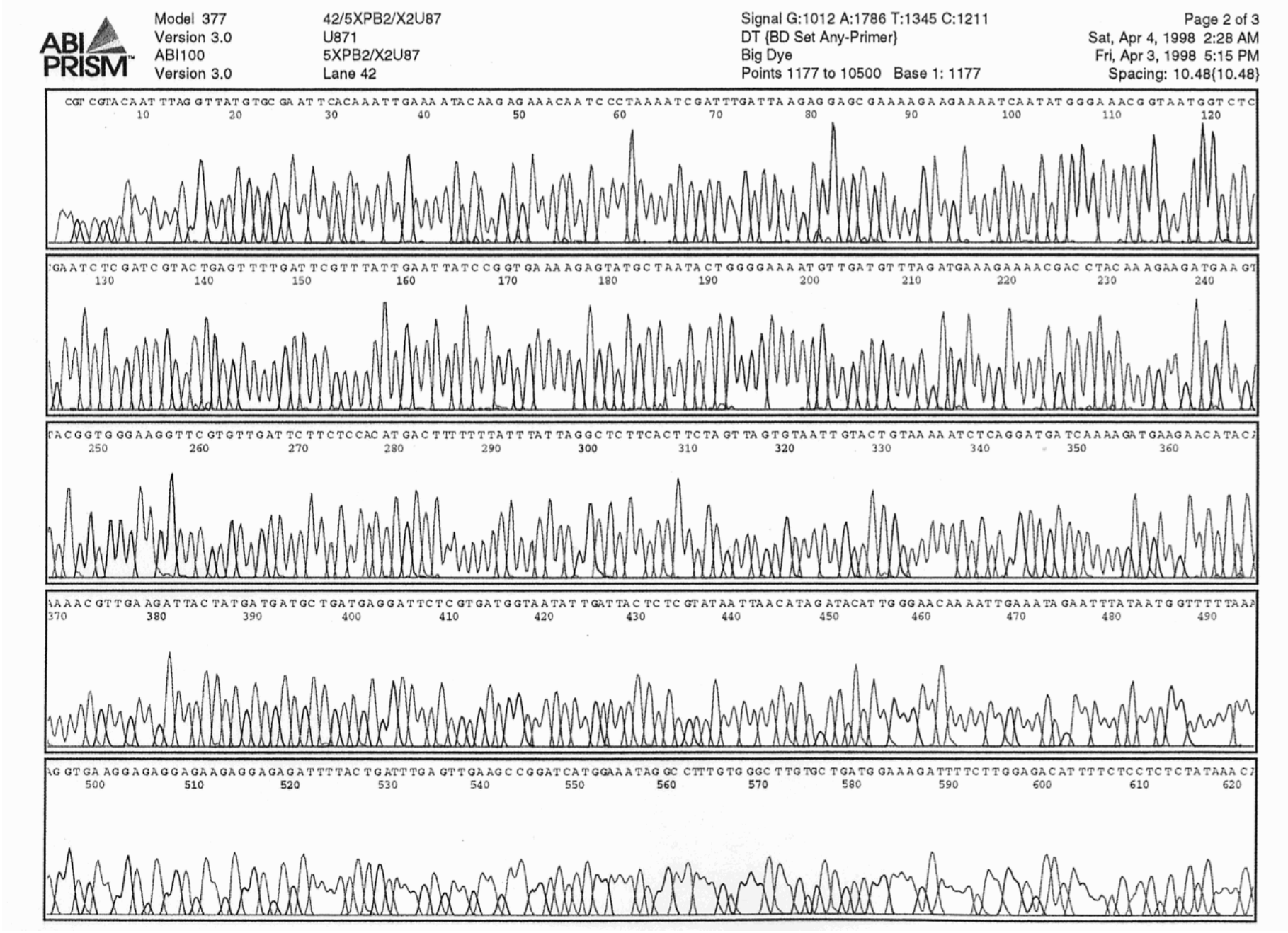
③ シークエンサーによるDNA断片の配列（600～700塩基）の決定

④ コンピュータによるDNA断片のアセンブル

DNAの配列

AAGAACCAGTTTGTAACCTCAAGAAGTGGGAAAATAACACTAACTTTATATTGGAAGATCATACGAAGAATAAAACAAAACTTTCTAGCACAGATAGGATACGTAAGTGGTTTAGAAGACATATATTAAAGGAAGAGATCGAAATCCTTTCCCATGGAAAACAATTGAGTAGTATTGAT
GAGGATTATTGCCCTTCAAATGTTCTTGTAGGATGTTCAAGAGATCTAAATAAACTCAGATCATTTCAAAATTTTTAGGAATATAAAATTTCTCAGTAAGAAAAAGAAGATGAATGCGTTTATAGTATTTTAAGTTTGTTCAATAGTTCATATGTGGACTAAAAG
AAGATTTTGGTATATGTCTATAAATTTTATTGTTGGTAAACGCCATTAAGCAGCTTTCCAAGAATATGTTATCATTGAGGAGAAGTACGAAAACTGATTATGCACGATTAGGAAAAATTCTGCAAGGAGTATTAAAAAATGGCCTTCGAGAAGAAATGAAATCTCTTTCGATGAGCAATG
GTCTTATGGATAATTTGATAGACATATGAAAAAGTCAAAAGTTTTTTTGGCCATATGCAGTAAAAACTTACCCTTCAAATTGCAAAACCATTAGTTTTATTATATTTTTTTTTTTCTTTCATTGAAGTATACACGAAATCCCATACGCAAATAAACAGTCGTTGTATCATCAATTTC
GGCTGTCTGGATGGCGACTACGAGAGGAAGCTTGCTTGCCATGATAGGACCAATCTCAATTTTTAATTTTTCATTAACTAGCTAATTTTCTATAGTTCTCTAAGTTTCCTTTTTCGTTTTGTTTCTTTATAGTGATGATGCTGTTATTGTGACGCTATCAAATCATATACTGTCATCAA
TCCTCATCTCATCGGCTCCGGAACGCGTTGAATATGTTGTTAAACAAGCTCTTTATAGCTAGCTATAGTAGAAAAGGAAGGAAAGAAGTCTAATCGCTGTATTATCGACTGACAAGGGTTGTCAGTTGCCTTCTTTTATTTTTTGTACTCCCCTGAATCGAGCGGATGTATTTTTTAAC
AAAGTATTCAGAATTGTTTTTGTATAATCAAAATTTGTCGACGCTAAGAAAATGACGGACGTCGTAATTCCGCTACTAGGTTTTGTGTGCTCATCGATATTTGACACCTATTTTCGACACTTTTAAGGTTATATACTTTGTTTCGGATGTGGGGTACACTCAGTTGACTGATTATGAAA
GATATGGCGTAGCCGTTGGGTTAGGGTTATTACACAGAAATGAATTATACCAGTGAAATTATTCTCTTTATGTTTGTTCCTACATGTTTTCCAATCAGAAGGAATTGTTATATAAGGACACAAGAAATGTGTTGGTACGCTCTGCGCCTCAATGACGGTTTATCTATCTAAAACTATTT
CTTTCTCGTAATAAAAATACAATAAATAATTAATAATAATATAGTTGATCGATATAACAAAAAAATAACAGCGAGGATGGTTAGTTCAAGTGTTTCCATTTTGGGGACTAGCGCCAAGGCATCCACTTCTCTAAGTAGAAAGGATGAAATTAAACTAACCCCTGAAACAAGGGAAGCTA
GCTTGGACATTCCATACAAACCCATTATTGCATACTGGACGGTGATGGGTCTCTGTCTGATGATTGCCTTTGGTGGATTCATTTTTGGTTGGGATACAGGAACCATTTCAGGGTTTATTAACCAAACAGATTTCGGTGAGTTACAAAGGGACGGCAGTTTTCAACTA
TCAGATGTCAGGACAGGGCTAATTGTCGGTATCTTCAACATAGGTTGTTGCTTTAGGTGGCCTAACGCTGGGACGCCTGGGCGATATTTATGGGCGTAAAATCGGCTTAATGTGTGTTATACTGGTGTATGTTGTTGGTATCGTGATCCAGATTGCTTCCTCTGACAAATGGTATCAATA
TTTTATTGGTAGAATTGTTTCTGGAATGGGTGTTGGAGGTGTTGCTGTGCTGTCGCCAACTTTGATCTCAGAAATTTCCCCAAAGCACCTAAGAGGCACTTGTGTCTCTTTTTACCAGCTAATGATTACCCTTGGAATTTTCTTGGGCTACTGTACCAATTATGGTACAAAGAAATATT
CAAATTCAATACAGTGGCGGGTTCCCTTGGGTTTGTGTTTTGCGTGGGCAATCTTTATGGTGATTGGAATGGTTATGGTTCCGGAATCGCCCAGATATTTAGTAGAAAAGGTAAGTATGAAGAAGCTAGAAGGTCTTTGGCCAAATCAAACAAGGTCACAGTTACTGATCCAGGCGTT
GTTTTTGAGTTTGATACTATAGTTGCAAATATGGAATTAGAAAGGGCTGTTGGAAATGCCAGTTGGCACGAACTCTTCTCAAATAAAGGAGCAATTCTACCAAGGGTAATAATGGGAATCGTTATCCAGTCACTGCAACAGCTTACTGGCTGTAATTATTTTTTCTACTACGGCACGAC
CATTTTCAATGCTGTTGGAATGCAAGACTCTTTCGAGACTTCCATTGTCCTTGGGGCTGTTAATTTTGCTTCTACATTTGTTGCACTATACATTGTGGATAAATTTGGGCGTCGAAAATGTTTATTGTGGGGGTCTGCCTCGATGGCAATTTGTTTCGTCATATTCGCCACCGTTGGCG
TCACTAGATTATGGCCACAAGGGAAAGACCAACCTTCTTCGCAAAGTGCTGGTAATGTTATGATCGTTTTTACTTGTTTCTTCATTTTCTCTTTTGCCATTACTTGGGCTCCTATCGCCTATGTCATTGTGGCAGAAACTTATCCATTAAGAGTTAAAAATCGTGCCATGGCCATTGCG
GTTGGTGCGAACTGGATGTGGGGTTTCTTGATTGGATTTTTCACACCCTTTATCACTAGATCCATAGGATTTTCTTATGGCTATGTTTTCATGGGTTGCTTTAATCTTTTCGTACTTCTACGTTTTCTTCTTTGTTTGCGAAACAAAGGGATTAACTCTGGAGGAAGTTAATGAAATGTA
CGAAGAAAGAATAAAGCCATGGAAGTCCGGAGGTTGGATTCCCAGTTCTAGAAGAACACCACAACCAACAAGCAGTACACCATTAGTTATTGTTGATAGTAAATAATTTCTAAATATTCTTGTACTCTGGTAAACAGAAATAACAACAGATAATGGATTGATGCTTTACTTCTATTTCA
TGGAGATTGGTTTTATATAATAGCCTTTATTAATGGCGTCACAAATTGAAAAAAAAATTAAAAAAAATAAGACGGACACATTTGCGACGCCGGTGAATAAATGCATATAAGTAGTTTATAAGCTAGCTACTACTCAGAAATAATTCTCAAAAAAAAAAGAAGTGGGCACTTTAAAATGA
AGATTAGCTATAGTAAATTACTGTAGTATAACAACTACCTGCTATCTTTCTGAAAAAATCAGGCAATGTAAGGTCAAACTTGTACCATCGACATATATAAGTGTTTGAGATATAAGTAACTAGAGACCAGTTTATACAGGATCTTACCTCTTTTTACCGTTATGAAAGCTTTATTACTG
CGTTGTTTAGTAAACACACATATTATTTTCGGGTAGTCCTGTCGATGTTCCAAGGCTTTGCATTTATCATTGTTTCCGCTTCATCGGTGGTATTAGTTAGTGTATCTGTATGTGAAGTATGTGGATAGGTGCTTCTATTATTGGAACAAAACACCTTAAACGCGCACTAGGTTATAGGAA
AGGGTCCATTATTCAAAGAACGGCTTATTGAAAAGTATGTTGTAAAGCTCGGTCATATCGCGCATAGCCACGATAAGGCGTGGTGCTGCTTCATAGATTGAGAGGCGCAGTCAATACACAACCATACCGGTAGTTTGATAAATGATACTATCATTCCGGAAGCTCTCTAGTAAGCTGTA
GGTGGCATATACGGTATCTATCATCTAGTAATAGCTATTGATTTTTTTCTCGACTATGCTATATGTGTTGTGATGGTTCACTTAGAAGTGAAAAAGTAGCAATAAATTAGACAGACAGAAGTAATTGAAGCTACATTCAACAACATGGCTAAGTAAAAAAGCGTAGAATATCATGATGG
GTTTTTTTATCTTTTAATTGCTTTCTATGTAGTATAATGGACCGATCTTTGCAAGTATATATCTGTATGTATCCATATTTAGATGGCAGCAAGCAATATAGATTTGATGAGCTTATATCATTTTATCGTCCTTGTCCAAAAAGTCTTGATAACATTAAAAGTCACTACCGTCAAATCCA
TCATCAAATCCGCCGTCGAACCCACCAGCATCATCAAATCCGCCGTCGGACCCACCAGCATCATCACCGTAGTAATTGTTCTCGACAACGACAGTGTCTGGTCCGTCATAGTTGTGGTCGTCAAATGCGTGTTCTAGCATAGCTCCACCTAACAAACCCGCACCAACACCAAGTAAACC
ACCCATCATGGCACCGCCGTGTCCACTACCTTTACTGGTGCTTGGCGCAGTTCCATAGTATGCTTGCTGTGGTGCTGCTGCTGGCATAGGGGCTTGTTGAGGATAGTAACGCTGCTGTTGAGGGTATTGAGGGTACTGAGGCTGTTGAGGCTGATAGTAACGAGGTTGCTGTGCTTGGG
CCCCTGCTTGCACTTGCGGTGTTGATTGAGAAGAATAAGGTGGAGGAGCCTGGTCTGCCTGTTGACGAGAACTCTTCTCATTGTTAACACCTGGTGGAGGACCTTTGGGTCTTGGCCATGTTGTTCCCCTTGGTGGTTCCCACTGAGAGCTGTTCGTAGATAAATCTACATAATACCAA
GTCTGGTATTCATCATCAAAAACAGCCTTCCAGCCAGAAGGCACTTGAGGAGGATTACTTTTACTTTGAGCCATTACTTTAAAATGTTTTGGGTTTTTTTATCGACTCTCTTGTAGTGAGTGATAAAAGCAATTCAGAAAGAGTCTTTTGATTATCATGACTAAAGTAAACTTTTTCTT
CCAGCTGCCGGCTCGGTCAGGTGGACGAAACAAAAAAGCCTGTTGAAATAATACTGAATACGTTAAAACGTCACAGCGCTTGTATATCTTGTGAGAGAAAGGCTATTACCACTGTTATATGAATATAGATTGCTCAACAAAGCGTGTCTTTTTCTGTGTGCTTAAAATATAGTGTCTTT
TTTGGGAAAAAAACGTTTATTTACAGTATTCTCTTCTTCTTCTCCTCATTCAACTTGTTGCGGTTCGAAGAAATTACGCATAAAGAATCAAAGAAAGGCAAAAATTACGCTGTACGATGGGGTCGTTTCCCTTAGCTGAGTTTCCATTACGTGATATCCCTGTTCCTTATAGCTACCG
TGTGTCTGGCGGTATAGCTTCCTCAGGTAGTGTTACTGCGCTTGTTACTGCCGCTGGCACTCATCGAAACTCGTCCACGGCTAAGACAGTTGAGACAGAAGACGGCGAAGAAGATATCGATGAGTATCAGAGGAAAAGAGCAGCTGGTTCTGGCGAATCCACTCCTGAACGCAGTGATT
TCAAAAGGGTAAAACATGATAATCACAAAACCCTCCATCCAGTTAACTTACAGAACACCGGTGCAGCGTCTGTGGATAATGACGGTCTGCACAATTTAACAGATATATCCAACGATGCAGAAAAACTTTTGATGTCTGTGGATGATGGTTCTGCCGCACCTTCTACATTGAGTGTAAAC
ATGGGAGTGGCATCTCATAATGTTGCTGCTCCCACTACCGTCAATGCGGCAACAATAACTGGCAGTGATGTTAGTAACAATGTTAATAGTGCTACTATTAACAATCCTATGGAGGAAGGAGCGCTGCCGTTTCCTCTCCAGGTACCACAACTCCTTTAGCTAAAAC
TACGAAAACTATCAACAACAATAATAATATCGCCGATTTGATAGAATCCAAAGATTCTATAATCTCCCCTGAATACCTTTCTGATGAGATTTTCAGCGCAATAAACAATAATCTCCCTCACGCATACTTCAAAAATTTATTATTTAGATTAGTTGCCAACATGGATAGGAGTGAACTAT
CCGACTTGGGGACTTTAATCAAGGATAATTTAAAGAGGGACCTAATAACGTCTTTGCCTTTTGAAATAAGTTTGAAAATTTTCAATTATTTGCAATTCGAGGATATTATAAATTCCCTTGGGGTCTCCCAAAATTGGAACAAAATAATTAGAAAATCTACATCGTTGTGGAAAAAACTT
CTGATATCGGAAAATTTTGTGAGCCCAAAGGGTTTTAATTCTCTCAATCTCAAACTCTCCCAAAAATACCCAAAACTCTCACAACAAGATCGCCTTAGATTATCTTTTCTGGAGAATATATTCATTTTAAAAAATTGGTACAATCCCAAGTTTGTACCACAAAGGACCACGTTAAGAGG
CCATATGACGAGTGTTATTACGTGCTTGCAATTTGAAGATAATTATGTCATTACGGGGGCTGATGACAAAATGATCAGAGTTTATGATTCGATAAACAAGAAATTTCTTCTACAACTATCAGGTCATGATGGTGGGGTTTGGGCATTGAAGTATGCCCATGGCGGTATTTTAGTCAGCG
GTTCTACAGACAGAACGGTGCGAGTTTGGGATATTAAGAAAGGTTGTTGTACCCATGTGTTTAAAGGTCATAACTCTACGGTGAGGTGCCTAGATATAGTAGAATATAAAAATATCAAGTACATTGTTACTGGTTCGAGAGATAACACTTTGCACGTTTGGAAATTGCCCAAGGAGTCC
TCCGTTCCTGATCATGGGGAAGAACATGATTATCCATTAGTCTTTCATACCCCTGAGGAGAACCCATATTTTGTTGGTGTTTTAAGAGGACATATTGGCATCTGTAAGAACTGTCTCAGGCCACGGTAAATATTGTCGTTAGTGGCTCCTATGATAATACACTGATTGTGTGGGATGTTGC
GCAAATGAAATGTTTGTATATTTTAAGTGGACATACGGATCGTATTTATTCGACAATCTACGATCATGAAAGAAAAAGGTGCATCTCTGCCAGTATGGATACCACTATTAGAATTTGGGATTTGGAAAATATATGGAATAATGGAGAATGTTCCTACGCAACAAATTCAGCATCGCCAT
GCGCCAAAATACTTGGTGCTATGTACACTTTGCAGGGTCATACAGCTTTGGTCGGTTTATTAAGATTATCCGACAAATTTTTGGTCAGTGCCGCTGCAGACGGTTCAATAAGGGGTTGGGACGCAAACGACTACTCTAGAAAATTTTCCTACCATCATACCAATTTGAGTGCAATTACC
ACATTTTATGTATCGGATAATATTTTGGTGAGTGGATCGGAAAATCAGTTCAACATCTATAATCTACGGAGTGGGAAATTGGTCCACGCAAATATTCTAAAAGATGCTGATCAGATTTGGTCGGTTAATTTTAAGGGCAAAACACTTGTTGCAGCAGTTGAAAAGATGGACAGAGCTT
TTTAGAAATTCTGGATTTCAGCAAAGCTTCAAAAATTAACTACGTTAGCAATCCCGTAAACTCCTCGTCGTCGTCTTTGGAATCCATTTCTACTTCTTTGGGTCTAACGAGGACAACTATAATACCATGACCTTTCCCAGAGAATAAGCATTGACTCATACTTAGATAATATAGCTTAA
TAAGTAGTTATATAATCAGTAAAAAAGTACAATAACAACTTCGTACATTTTATTGAGTATAAACTGCAGCTAAACTGCCTGGATGTGTCAATTTTAATTGTGTTTACAAAAAGGGTGCCGTTTATTAATTAATGTTTCTTCCCTGAAAATATGGAAAGTACAAGTTTTTAGTTGAGAAG
GGTTTAAGAAAGTTTGAAAATGATCTAAAAAAATATAAAAGCAATCAAAGAAATAAAAGCTGGAAAAATGCGTAATAACCGAAGTGACTAAAATTTCTTTACGCGCCAAATAAGAAATCGATGCTCTTCGTATGGTTCTTGGCCC
AAGAGTTTTCTTGAACTTTTAACGTTAAGAAGTTGATTCTGCTGATTTTTTTCAAGTTATCAAGCGTTATGTTTTGAAACATCTGTTCCGTTTTCAGTTTTTCAAAAAGATCGATATTTCTCTTGAGAATTAGGAAGTTCTACGCTCAACCTCCCGTGGGCAGAAGAAATACAAAAGCTA
ATACAATTGTGTTAGAATAAAGTTCTAATATTATCTAATTAGTAGTATTCATGTTACTAGTATATTATCACATGATATTTCTCAAATTGGACAATGTAAATAATGAGTGTTTTTATGACACAATCTTTATCGTTAGATGTTTCAACGTTCCAAGGCTTGGTTCTATCGCTCTTCTCTTC
AAATTGTAATCGTTTGTCATGATAAACACGTACGGGAAAAAAAAAAAAAAGTATCAATAACGCGTAAAGTGAATAGAGTATTGGATTCTATAAGACCGAAGACGCTCATATCACATCTCATAAAATCACTTAAAGCAAGCATCCAGAGGCTATTGATAAAAAGCAGGCACAAGGAGACGC
AATGGGACGTTTAGTTGGCTTAGAACTAAGTAATTTCAAGTCCTATAGAGGCGTTACCAAGGTAGGATTCGGCGAGTCGAATTTCACAAGTATTATCGGTCCTAACGGATCTGGTAAATCGAATATGATGGATGCTATCTCATTCGTACTCGGTGTGCGGAGTAATCATTTGAGGTCAA
ACATCTTGAAAGATTTAATCTATAGAGGTGTTCTAAACGATGAGAATAGTGACGATTATGATAACGAGGGCGCTGCCTCTTCGAACCCACAATCCGCATACGTGAAGGCATTTTATCAAAAAGGTAACAAACTGGTGGAGCTGATGAGGATAATTTCCAGAAACGGTGACACTAGTTAT
AAAATTGATGGAAAAACTGTCTCCTATAAGGACTATTCTATATTTCTTGAGAACGAAAATATTCTTATCAAAGCCAAAAATTTTCTAGTGTTCCAGGGTGATGTTGAGCAAATTGCAGCACAATCTCCCGTAGAATTATCAAGAATGTTTGAAGAAGTATCAGGTCTATTCCAATACAA
AAAGGAGTATGAAGAGTTGAAGGAAAAGATTGAGAAATTAAGCAAATCTGCAACCGAATCTATAAAAAATAGGAGAAGGATCCATGGAGAATTGAAGACATATAAAGAAGGTATCAATAAGAACGAGGAGTATAGGAAACAATTGGACAAAAAAAATGAATTACAGAAGTTCCAGGCTC
TATGGCAGTTATATCATTTAGAGCAACCAAAAAGAGGAGCTAACGGACAAGCTGTCCGCATTAAACTCTGAAAATATCGTCTTTAAAGGGAAAAATAAATAACGAGATGAAATCATTACAACGCTCAAAATCTTCCTTTGTTAAAGAAAGCGCAGTAATTCTTAAGCAAAAAAGTAAATTA
GATTATATCTTCAAGGATAAGGAAAAATTAGTCTCGGATTTACGGCTCATAAAGGTTCCTCAACAGGCAGCAGGGAAGCGAATTTCCCATATAGAAAAAAGAATCGAAAGTTTACAGAAAGATCTTATGTGGAGAGATTTGAAACACAACTAAAAGTGGTGAC
CAGATCAAAGGAAGCTTTTGAAGAGGAAATCAAACAATCTGCTAGAAACTATGACAAATTCAAGCTAAATGAAAATGATTTAAAGACATATAATTGCTTACATGAAAAATATCTGACTGAAGGTGGGTCAATCCTAGAAGAAAAAATTGCCGTTTTGAACAACGATAAGCGAGAAATCC
AAGAGGAATTGGAGAGATTCAACAAAAAGGGCAGATATTTCTAAAAGAAGGATAACGGAGGAGCTTTCTATAACAGGAGAAAAATTGGACACGCAATTAAACGATTTAAGAGTTTCTTTGAATGAGAAAAACGCCCTTCATACTGAACGTTTGCACGAGCTGAAAAAATTACAATCTGAT
ATTGAATCTGCTAATAATCAAGAATACGACTTAAATTTCAAGTTGAGGGAAACGTTGGTTAAGATCGATGACTTGAGTGCTAATCAAAGAGAAACAATGAAAGAAAGAAAACTAAGAGAAAATATAGCAATGTTGAAAGATTCTTCCCCGGAGTAAAAGGTCTTGTTCATGATCTTTG
TCACCCAAAAAAGGAGAAATATGGCTTGGCAGTGTCTACCATCTTAGGTAAGAACTTTGATTCCGTCATTGTAGAAAATTTAACCGTAGCTCAAGAATGCATTGCATTTTTGAAGAAGCAACGTGCGGGCACTGCATCTTTCATACCACTAGACACAATTGAGACAGAGTTACCTACAT
TATCATTGCCTGACTCACAAGACTATATTTTATCAATTAATGCTATTGACTACGAGCCGGAATATGAAAAAGCGATGCAATATGTGTGTGGCGATTCCATCATATGTAATACGTTGAATATTGCTAAAGATCTGAAATGGAAAAAGGGCATAAGAGGCAAATTGGTTACAATTGAAGGT
GCTTTGATCCACAAGGCCGGTTTGATGACAGGTGGTATATCAGGAGATGCCAATAATAGGTGGGATAAAGAAGAATATCAAAGCTTAATGTCTTTAAAAGACAAATTACTAATCCAAATCGATGAACTTTCCAACGGTCAACGCTCTAATTCAATCAGAGCAAGAGAAGTTGAAAATAG
TGTTTCACTATTGAACTCGGACATAGCAAATTTGAGAACTCAAGTAACACAACAAAAACGCTCCTTGGATGAAAATCGTTTAGAGATTAAGTACCATAATGACTTGATAGAGAAAGAAATTCAACCGAAGATAACTGAACTAAAGAAGAAGCTAGATGATTTAGAAAATACTAAAGATA
ATTTAGTGAAAGAGAAGGAGGCTTTACAAAATAATATTTTCAAAGAATTCACTAGTAAAATTGGCTTTACAATCAAAGAATATGAAAATCATTCCGGTGAATTGATGAGACAACAATCTAAAGAATTACAGCAGTTACAAAAACAAATTTTGACCGTTGAAAATAAGTTGCAGTTTGAG
ACAGACAGACTAAGTACTACTCAAAGAAGATATGAAAAGGCGCAAAAGGATTTAGAGAATGCTCAAGTTGAAATGAAGTCTTTGGAAGAACAGGAATATGCAATAGAAATGAAAATCGGATCAATAGAGTCTAAATTGGAAGAACACAAAAATCACTTAGATGAGTTACAGAAGAAATT
TGTAACGAAGCAAAGTGAATTAAATTCCAGCGAAGATATTCTAGAGGACATGAACAGCAACTTACAAGTCTTAAAAAGGGAAAGAGACGGTATAAAGGAAGATATTGAAAAGTTTGATTTAGAGAGAGTAACAGCGTTAAAGAATTGTAAAATTTCTAATATAAATATACCTATATCAT
CTGAAACAACGATAGATGATTTACCAATATCTTCCACTGATAATGAAGCAATTACAATTTCCAACAGTATCGATATAAACTATAAAGGACTACCTAAAAAATACAAAGAAAACAATACCGATTCGGCAAGGAAGGAGCTGGAGCAGAAGATTCATGAAGTGGAGGAAATATTGAACGAG
TTGCAGCCCAATGCAAGAGCTTTGGAGAGATACGACGAGGCGGAAGGAAGGTTTGAAGTGATTAATAACGAAACAGAACAACTAAAGGCCGAAGAAAAGAAAAATATTAAACCAGTTCCTAAAAATTAAGAAAAAAAGAAAGGAACTGTTCGAAAAGACATTTGATTATGTGAGCGACCA
TTTAGACGCAATCTACAGGGAACTTACTAAAAATCCCAACTCCAATGTGGAATTGGCCGGTGGTAACGCTTCTTTAACCATAGAAGACGAAGATGAACCGTTCAATGCGGGAATCAAATATCATGCCACTCCGCCTCTTAAAAGATCAAAGACATGGAATATCTTTCTGGTGGTGAAA
AAACCGTAGCTGCATTAGCTCTATTATTTGCTATTAATTCCTACCAGCCTAGTCCCTTCTTCGTGCTGGACGAAGTGGACGCAGCCCTAGACATTACTAACGTCCAGAGAATTGCTGCCTATATAAGAAGGCACCGTAATCCAGATCTCCAATTCATTGTCATTTCATTGAAGAACACC
ATGTTTGAAAAATCTGACGCTCTCGTTGGGGTTTACAGACAGCAACAAGAAAACTCGTCGAAGATCATAACTTTGGACTTGAGCAATTACGCAGAATAATCTATGAAACCAACCTCTGCTATAACCCGTCAAATAACTAATAATATCTATATAGGTCAACTAGCTAGTGCAATATCATA
GTAACAATAATATTAATAACGTCACTTTTTTTTCCAGGGTAACCCAATTGTGGTGGGTGGCGGCCGAGGTATCCCTTAGAAAAGAATTTTTTAAGTTCTTTCTCATCTCTTACCAGTGGAGAAGTACACGATATTTGCAAAGTCTGTCATCAGGGCTTGATAATAAAGCTGCATTAGATC

DNAシークエンサー

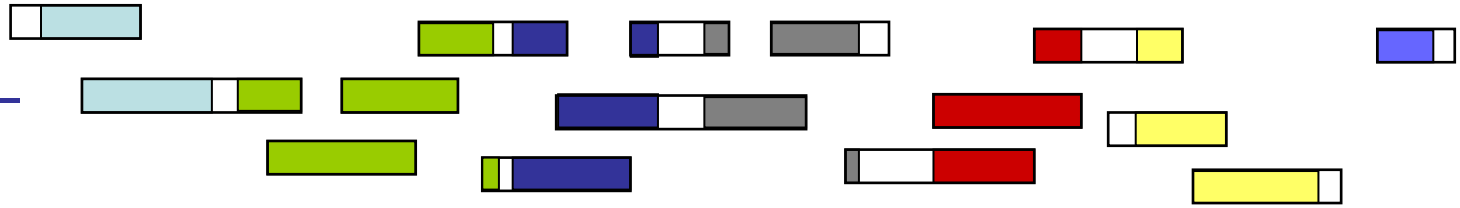

ABI PRISM
Model 377
Version 3.0
ABI100
Version 3.0
42/5XPB2/X2U87
U871
5XPB2/X2U87
Lane 42
Signal G:1012 A:1786 T:1345 C:1211
DT {BD Set Any-Primer}
Big Dye
Points 1177 to 10500 Base 1: 1177
Page 2 of 3
Sat, Apr 4, 1998 2:28 AM
Fri, Apr 3, 1998 5:15 PM
Spacing: 10.48{10.48}

ヒトゲノム配列決定のショットガン法

染色体:

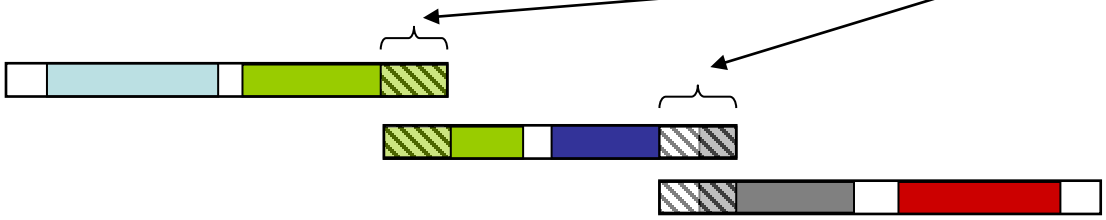
ショットガンにより断片化:

多重な
ライブラリー



シーケンシングにより配列決定⇒アセンブリ:

のりしろ



膨大な断片の両端を相互に比較して重なりを見つけて,
元の配列につなぎ戻していく:

ゲノム配列の決定法

① 大まかに分けて二種類考えられる

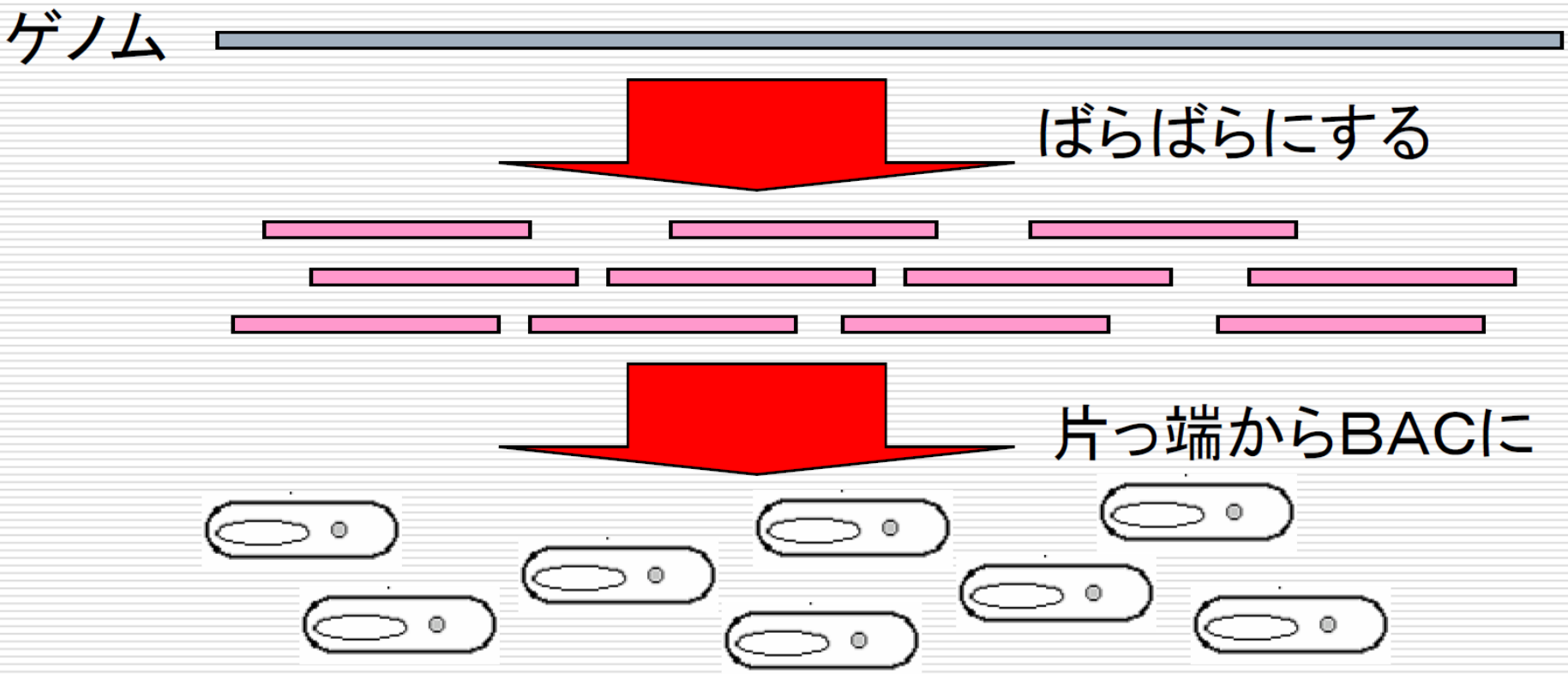
1. 階層的ショットガン法
2. ホールゲノムショットガン法

② それぞれ長所と短所がある

1. 階層的ショットガン法は，精度が高いが時間とお金と手間がかかる
2. ホールゲノムショットガン法は，お金や時間が少なくてすむが，精度の問題や長いゲノムの配列決定に対しては問題点も多い

階層的ショットガン法　①

□ ゲノムを小領域に分割し、BACに保存



サブクローニング（1/2）

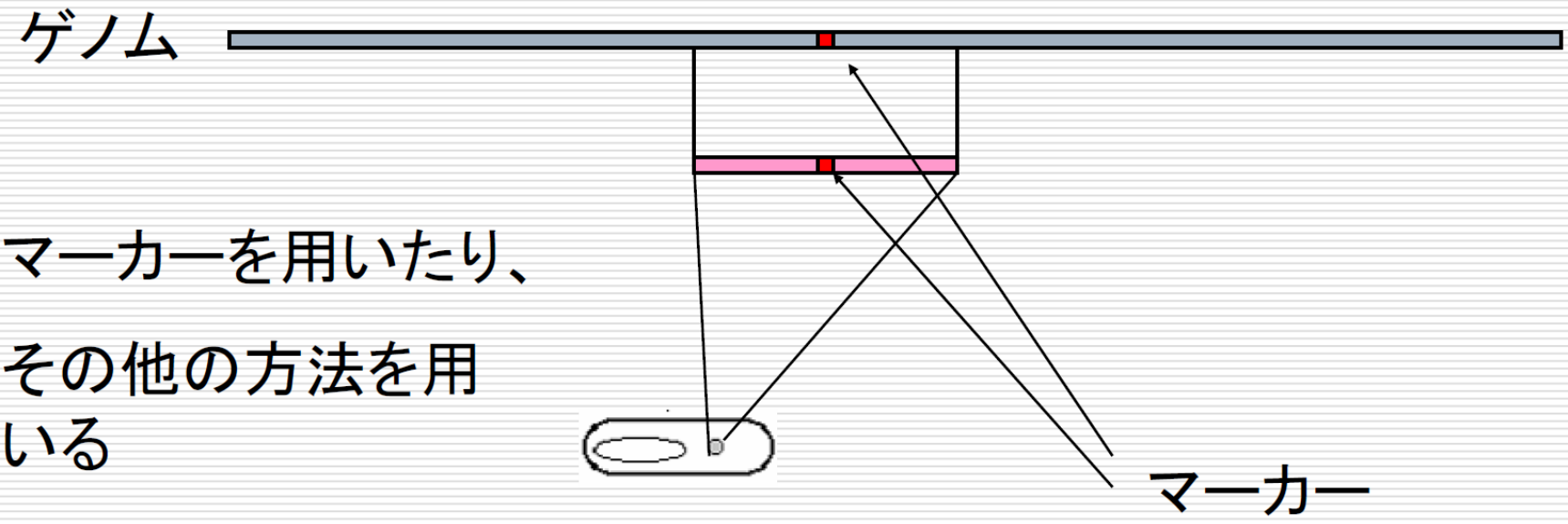
- 大腸菌や酵母菌の中に、人工的な染色体を注入すると、菌の増殖と共に人工染色体も増える→配列を保存しておくことができる
  - BAC（Bacterial Artificial Chromosome）
  - YAC（Yeast Artificial Chromosome）
  - …etc

サブクローニング（2/2）

- 人工染色体より規模の小さな配列を組み込むために、プラスミドやファージなども使われる（それぞれ組み込める配列のサイズが違う）
- 保持サイズの例
  - BAC (150kbp)
  - cosmid/fosmid (20kbp)
  - plasmid (2kbp)

階層的ショットガン法　②

- おのおののBACがゲノムのどこ由来のものか調べる

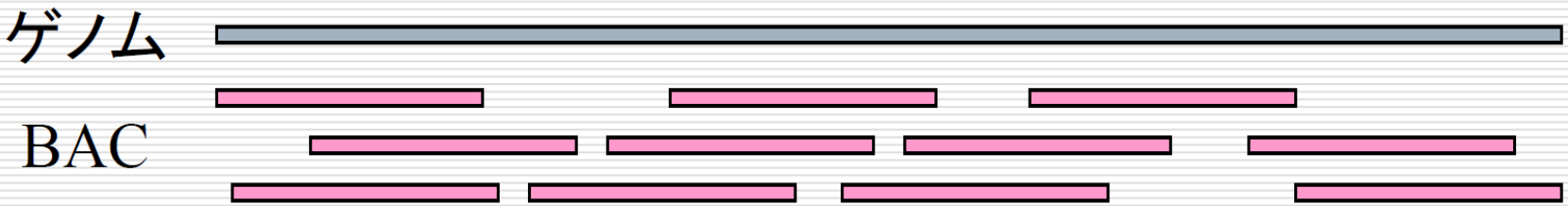


マーカーを用いたり、

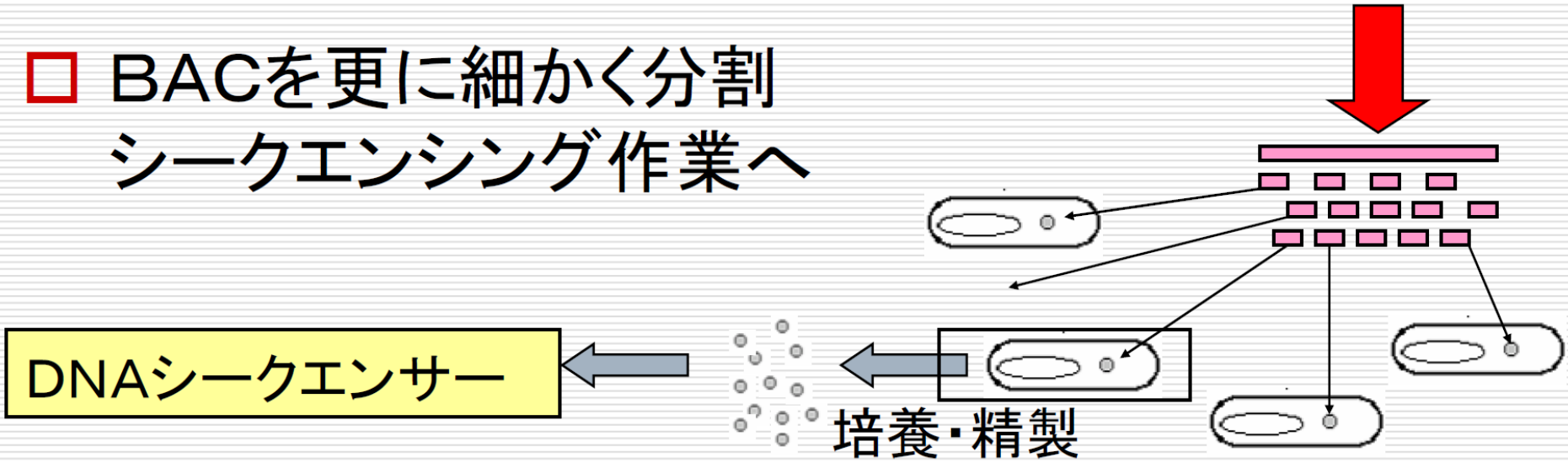
その他の方法を用いる

階層的ショットガン法　③

- ゲノムのカバー地図を作成する

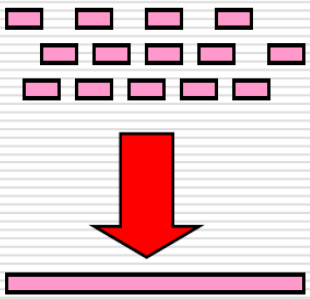


- BACを更に細かく分割
  シークエンシング作業へ

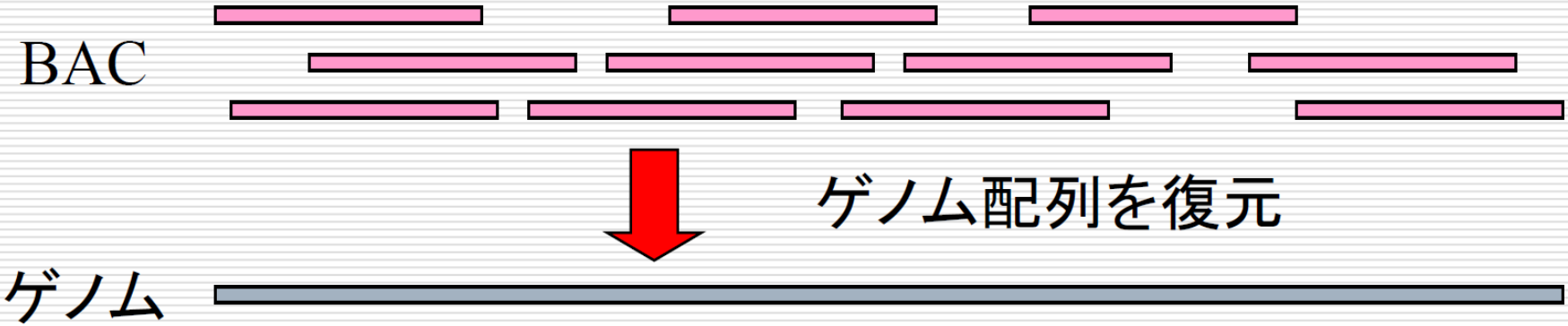


階層的ショットガン法 ④

- 得られたデータをコンピューターで接合し、オリジナルの配列を復元する。この作業を行うソフトウェアをゲノムアセンブラと呼ぶ



BAC配列を復元



階層的ショットガン法によるゲノムの配列決定の手順

*Not*Iの制限酵素サイト:

GC|GGCCGC
CGCCGG|CG

① ゲノムDNAを染色体ごとに分けた後，制限
～20万塩基対の断片に切断

② 断片をBAC（細菌人工染色体）にクローニング後，BACクローンのBACライブラリを作成

③ 物理地図を用いて，各BACクローンのゲノム（染色体）上での位置を決定

④ BACクローンの配列を決定するため，約2000塩基対の長さの断片，ショットガンクローン，に切断（超音波などにより切断）

⑤ ショットガンクローンの両端約600塩基の配列を多数決定し，それをコンピュータを用いてつなぎ合わせて，もとのBACクローンの配列を決定する ⇒ アセンブリ

⑥ BACクローンの配列と，整列順の情報を使って染色体の配列を決定する

階層的ショットガン法の特徴

① 精度の高いゲノム配列が決定できる

② 手間と時間とお金がかかる

- 大きなクローン（BAC）は作成に時間がかかる
- BACを並べ，マップを作るのが大変
- 手作業で並べなくてはならない

物理地図作成の方法

① 制限酵素マッピング：

制限酵素の切断部位をDNA分子に位置づける

② 蛍光 *in situ* ハイブリダイゼーション：

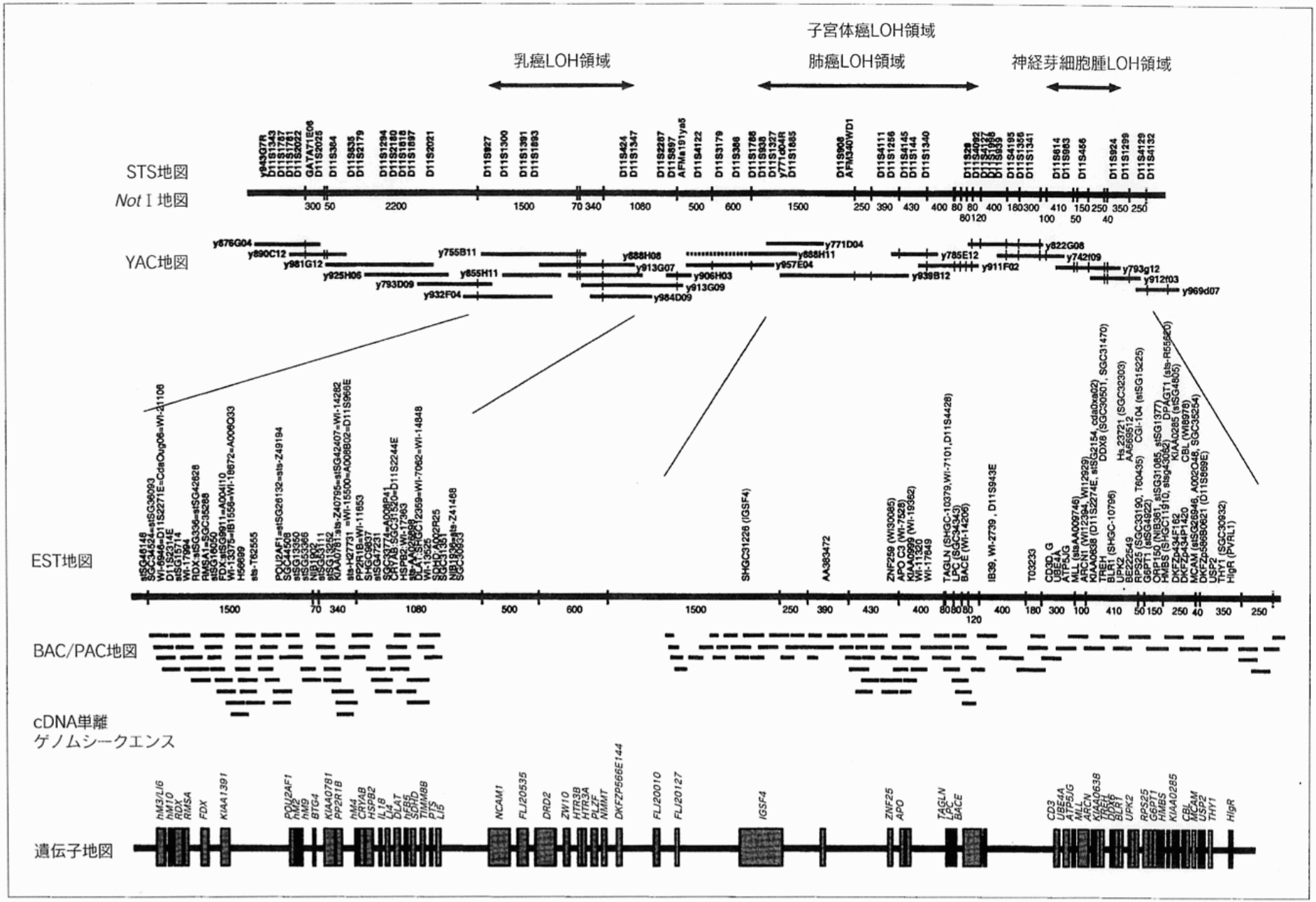
標識したDNA分子をプローブとして，無傷な染色体にハイブリダイゼーションさせることにより，そのマーカーの位置を決める

（DNAの標識：蛍光，放射性，その他のマーカーをDNA分子に付着させる）

③ 配列タグ部位（STS）マッピング：

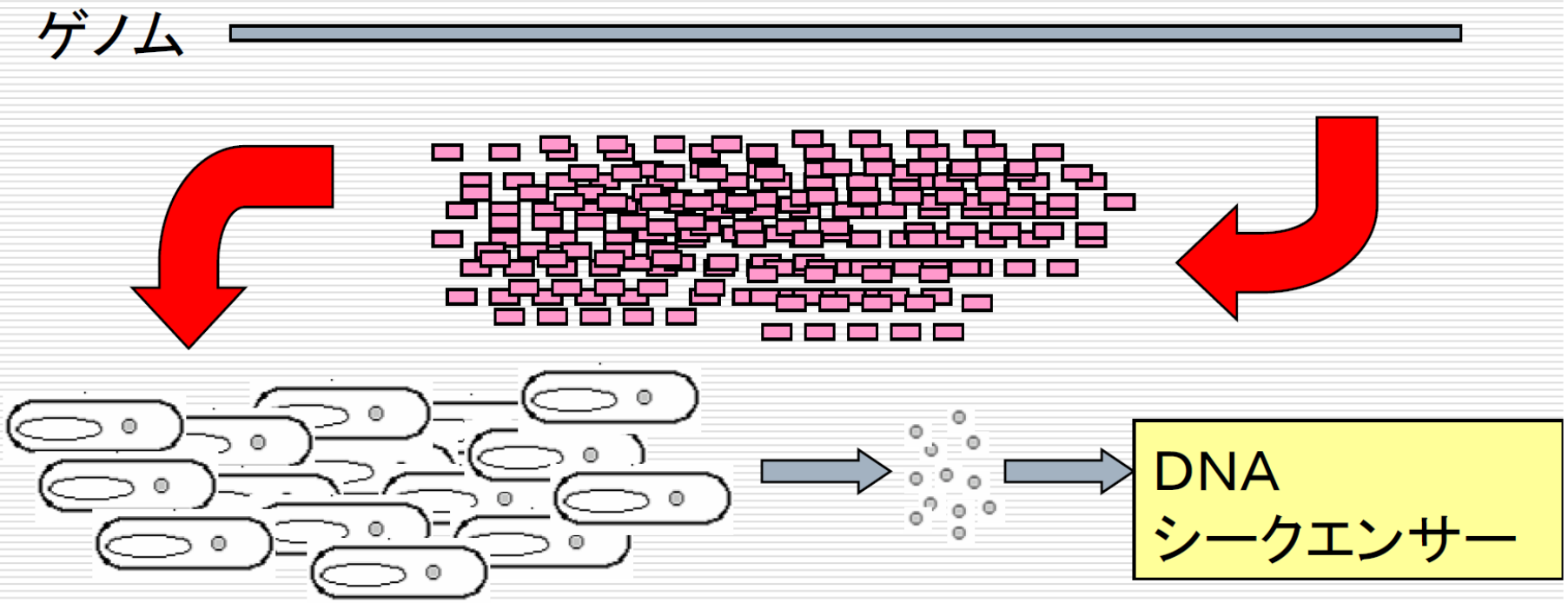
STSは100～500塩基対の短いDNA配列で，その塩基配列がわかっていて，対象の染色体やゲノム上にただ1度しか存在しないもの．このように定義されたSTSの位置を，多数のDNA断片の集合とPCRにより決定する

さまざまな物理地図


子宮体癌LOH領域
乳癌LOH領域
肺癌LOH領域
神経芽細胞腫LOH領域
STS地図
Not I 地図
YAC地図
EST地図
BAC/PAC地図
cDNA単離
ゲノムシークエンス
遺伝子地図

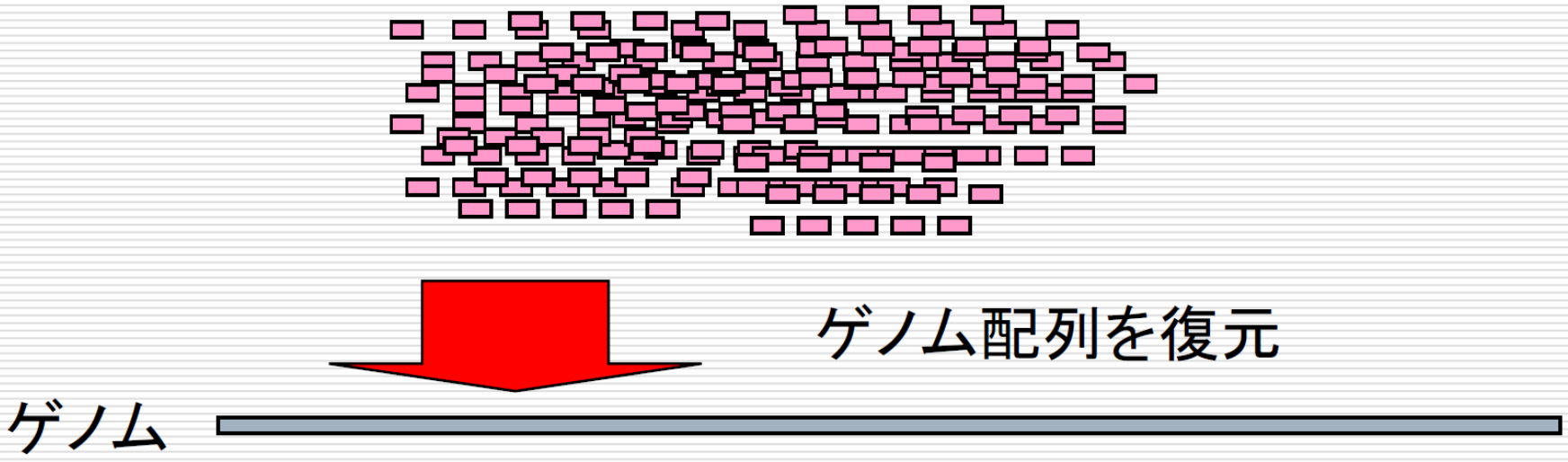
ホールゲノムショットガン法　①

- いきなりゲノムを細断片に分断、
  サブクローニングしてシーケンサーにかける
  （次世代シークエンサーでは，サブクローニングも行わない！）



ホールゲノムショットガン法 ②

- コンピューター処理により
  断片配列を貼りあわせてゲノム配列を復元



ドラフト配列

カバー率：10万塩基のBACクローンの配列決定を10万塩基分行った場合，カバー率1という．

シークエンサーの一度の決定塩基数を600とすると，10万塩基のカバー率1の場合には，約167(100,000/600)回のシークエンスをすることになる

- 確率論的に，正確な配列決定を行うには，10～20のカバー率が必要とされる → 完全配列
- ドラフト配列：全ゲノム配列に対して，カバー率4～5で配列を決定したもの．

  カバー率4～5では不完全で，7万5千ヶ所もの穴があいている，向きや順番が正しくないものもある
- 次世代シークエンサー（リード長数十塩基）の場合，ドラフト配列でカバー率40～50が必要．

コンピュータによるDNA断片のアセンブル

① 部分文字列をヒントにして，全文字列を決定する
（最短共通超文字列 （Shortest Common Superstring））

② ターゲット：再構築によって得られる長い配列

③ フラグメント：２重鎖のどちらかの鎖のある長さ分だけ
5’→3’の方向に配列決定したもの

④ 対象とする配列の長さを１０％程度以内の誤差で大まかに
知ることができる

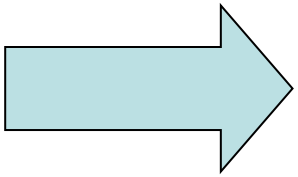
⑤ ショットガン法から得られた多くのフラグメント間の重なりを
もとにターゲット分子の塩基配列を再構築する

⑥ フラグメントをつなぎ合わせる作業を，アセンブルとよぶ

理想的なアセンブルの例

① ４つのフラグメント，ターゲットの長さ10塩基

ACCGT
CGTGC
TTAC
TACCGT



--ACCGT--
----CGTGC
TTAC-----
-TACCGT--
==========
TTACCGTGC

コンセンサス配列

配列の重複の情報を利用して，アライメント

実際のアセンブルにおける問題

① 部分文字列の誤り（ベースコールエラー）

② 部分文字列の向き

③ 繰り返し配列の存在

④ 全体が被覆されていない場合

部分文字列の誤り

① ベースコールエラー：塩基の置換，挿入，欠失

100塩基につき1から5塩基程度の頻度で起こる

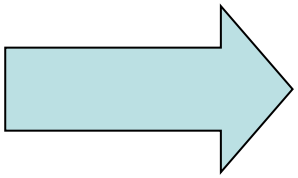
置換エラー

```
ACCGT          --ACCGT--
CGTGC          ----CGTGC
TTAC           TTAC-----
TGCCGT         -TGCCGT--
               =========
               TTACCGTGC
```

部分文字列の誤り

挿入エラー

```
ACCGT
CAGTGC
TTAC
TACCGT
```



```
--ACC-GT--
----CAGTGC
TTAC------
-TACC-GT--
==========
TTACC-GTGC
```

部分文字列の誤り

欠失エラー

```
 ACCGT                --ACCGT--
 CGTGC                ----CGTGC
 TTAC                 TTAC-----
 TACGT                -TA-CGT--
(TACCGT)              =========
                      TTACCGTGC
```

部分文字列の向き

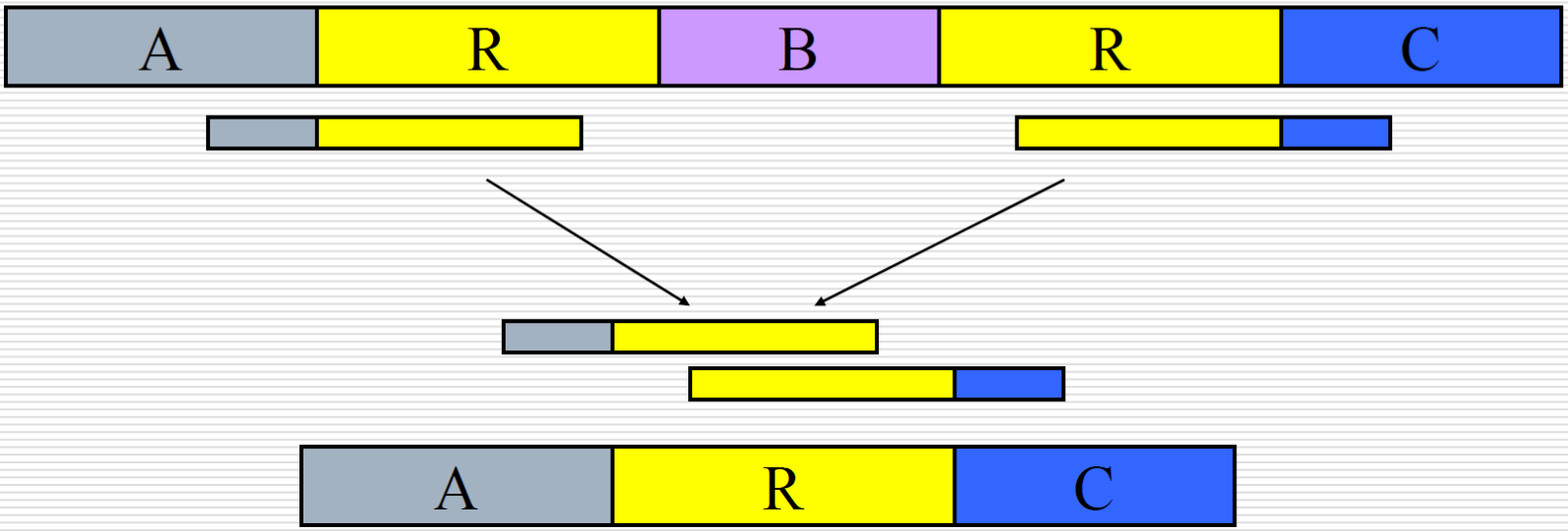
① フラグメントはDNA2重鎖の両方からくるので，通常はどちらの鎖に含まれるかわからない．相補鎖からの2つのフラグメントは，向きが反対でかつ塩基は相補的となる

```
CACGT      →  CACGT--------
ACGT       →  -ACGT--------
ACTACG     ←  --CGTAGT-----
GTACT      ←  -----AGTAC---
ACTGA      →  --------ACTGA
CTGA       →  ---------CTGA
           ==============
           CACGTAGTACTGA
```

繰り返し配列の存在

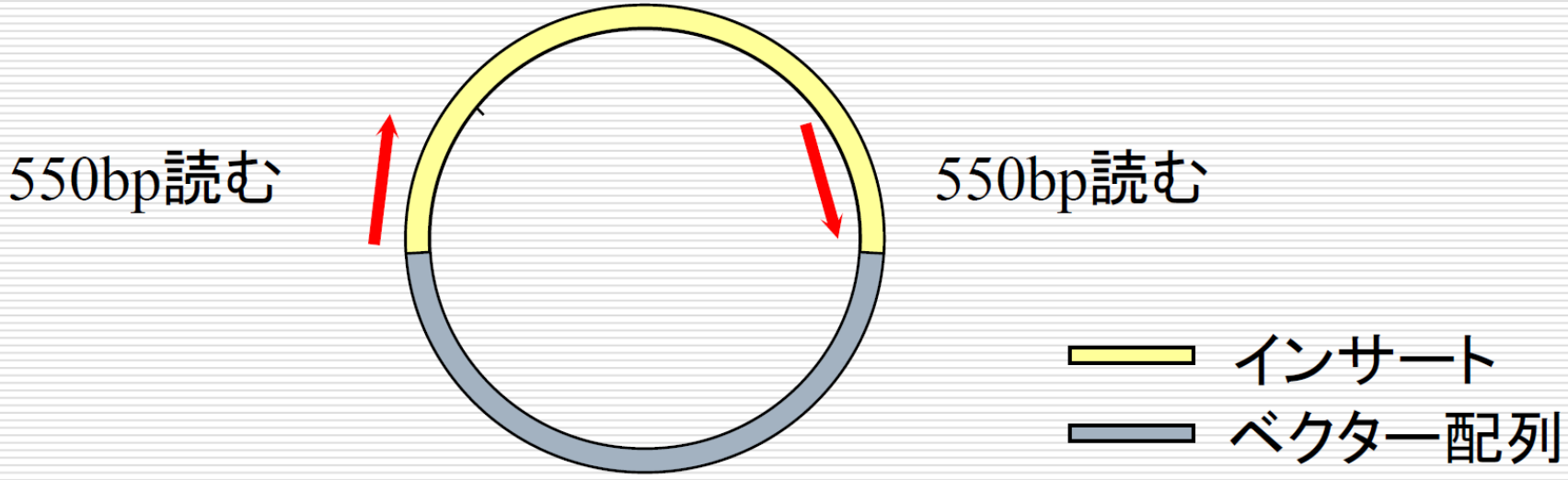
繰り返し配列：ターゲット分子内に，２回以上現れる配列

- ゲノムには進化の過程でduplicationが起き、大域的に同じ配列が存在することがある。
  - 一回に読める長さ以上のリピートRが来ると…



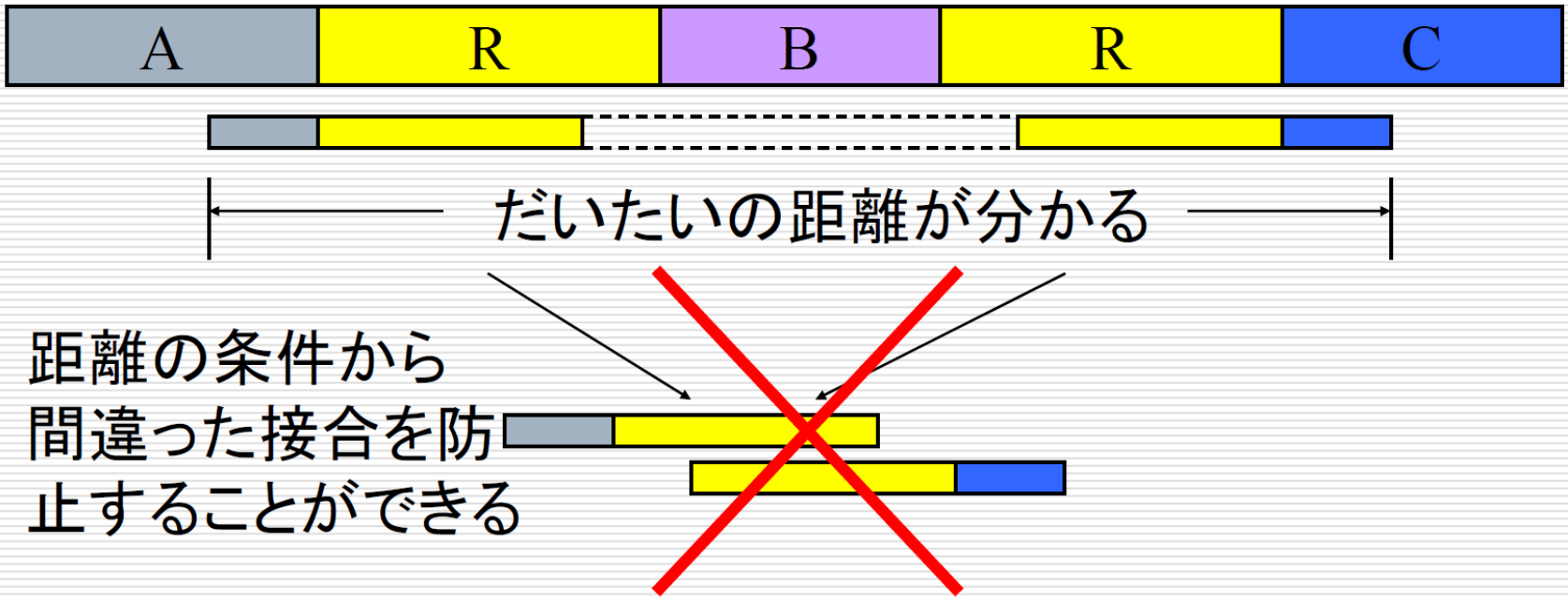
Double Barreled ショットガン法

- ベクターに組み込んだインサートの両端を読む



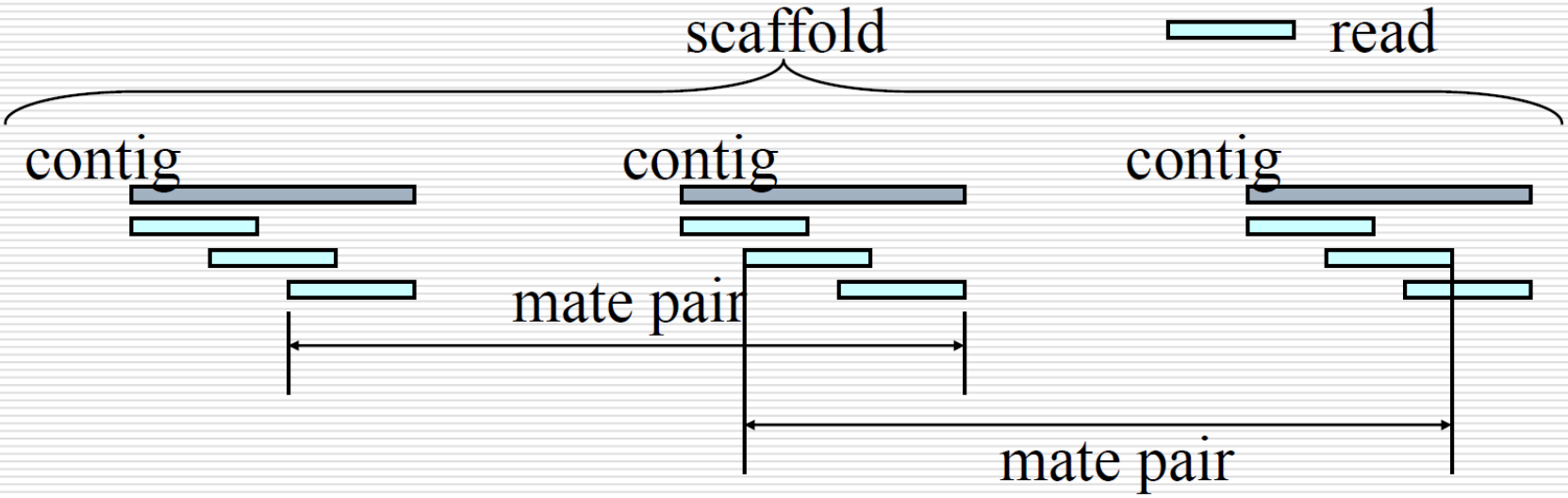
Mate Pair（ Double Barreled ショットガン法）

- 2つの配列と、その間のだいたいの距離が分かる。この2つの配列をMate Pairと呼ぶ



ゲノムアセンブラ用語

- Read（1回のシークエンシングで得られる配列）
- Contig（Contiguous sequence の略。Read を繋げ合わせて得られた連続配列）
- Scaffold（Contig を Mate pair で繋げた、隙間を含む配列群）



アセンブルのアルゴリズム

① エラーがなく，かつ向きが分かっていることを仮定

② 最短共通超文字列（Shortest Common Superstring）

③ ハミルトン経路アルゴリズム （グラフ問題）

最短共通超文字列（SCS）:

入力：文字列の集合 F

出力：次の条件を満たす最短の文字列 S:

S は，すべての $w \in F$ に対して，w の超文字列である

例） $F = \{$ACCGT, CGTGC, TTAC, TACCGT$\}$

$S =$ **TTACCGTGC** は F の最短共通超文字列

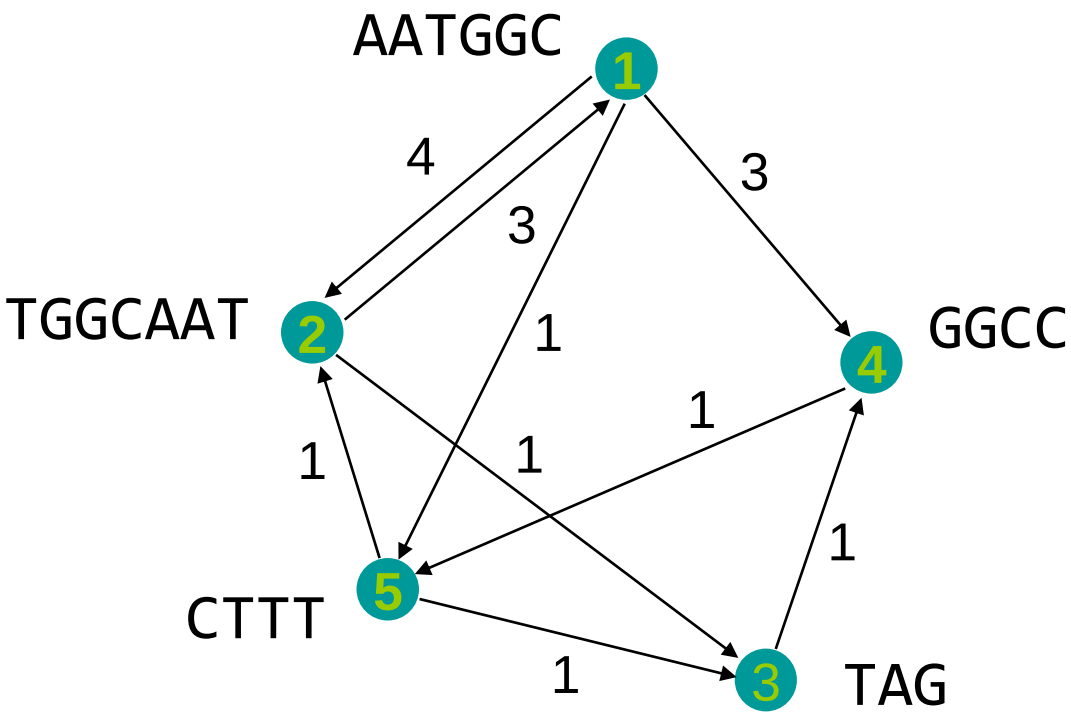
有向マルチグラフ

集合 F の有向マルチグラフとは:

① 各ノードには，集合 F の要素がラベル付けされている.

② ノード a からノード b に伸びる重み t の有向辺が存在する:

$$\mathrm{suffix}(a, t) = \mathrm{prefix}(b, t)$$



有向マルチグラフ上での最大重みをもつハミルトン経路は，
最短共通超文字列を与える

欲張りアルゴリズム

ハミルトン経路とは：

すべてのノードを「ただ一度だけ」通過する経路
（二度通過はダメ，一度も通らないのもダメ）

ハミルトン経路を求める問題は，計算量的に難しい（NP困難）

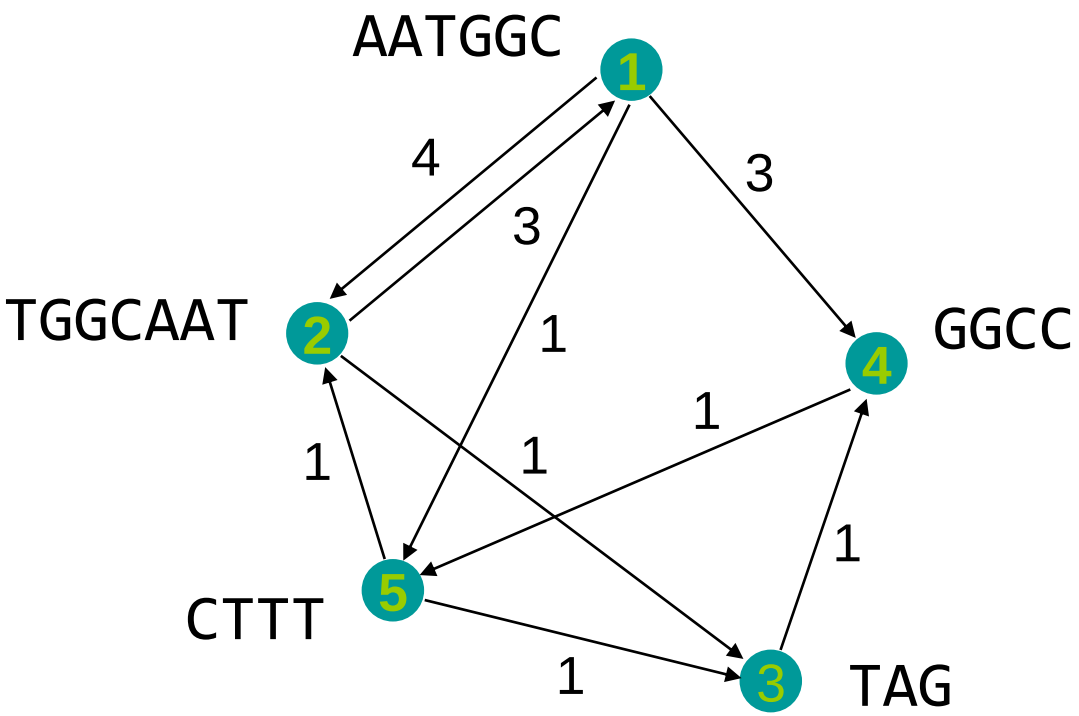
欲張りアルゴリズム：

それまでに選ばれた経路をそのままにして，ハミルトン経路の性質を乱さないように最大の重みの辺を追加していく

↓

必ずしも最大重みをもつハミルトン経路を見つけるわけではない

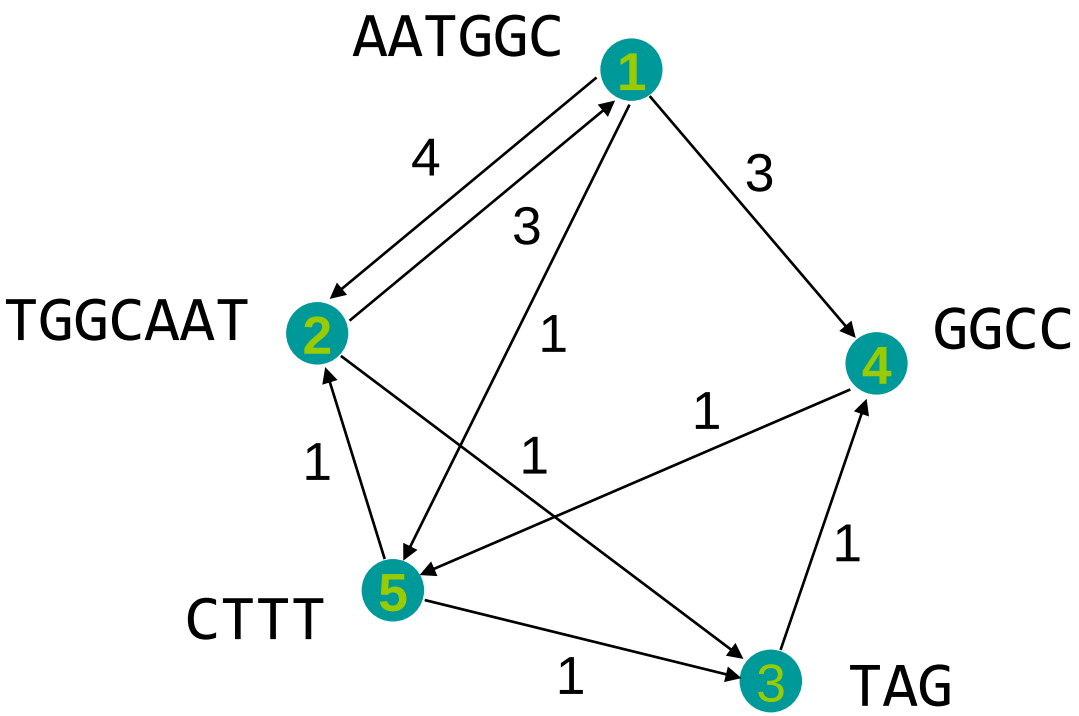
欲張りアルゴリズムの例



欲張りアルゴリズム：　1→2→3→4→5　重み：7

最大重みハミルトン経路：　2→1→4→5→3　重み：8

ハミルトン経路から算出されるアセンブリ



欲張りアルゴリズム：　1→2→3→4→5　重み：7

AATGGCAATAGGCCTTT（長さ17）

最大重みハミルトン経路：　2→1→4→5→3　重み：8

TGGCAATGGCCTTTAG　（長さ16）

ハミルトン経路から算出されるアセンブリ

最大重みハミルトン経路によるアセンブリ:

```
TGGCAAT---------
----AATGGC------
-------GGCC-----
----------CTTT--
--------------TAG
================
TGGCAATGGCCTTTAG
```

コンセンサス配列

アセンブリの演習

アセンブリの演習問題は，５月１０日（月）に対面で行う予定です