

第3編 健康科学・医療・創薬への応用

3.2 「さまざまな生命科学データの 接続で見える新たな知見」

生命科学研究における測定技術の進展は、膨大な生命科学データの蓄積をもたらした。これらのデータを利用して、新規の関係や仮説を生み出すためには、研究者の発想を支援するツールが必要である。そこで、生命科学の新たな知見（仮説）をデータに基づいて見いだすことを実現できるプラットフォームの構築に挑戦している。簡単なキーワードを入力することで、異なるデータベースを同時検索し、それらデータの関係を見いだすことができることをめざしている。

由良 敬

お茶の水女子大学／早稲田大学

2018年12月19日
17:00～18:30

「計算生命科学の基礎Ⅴ」

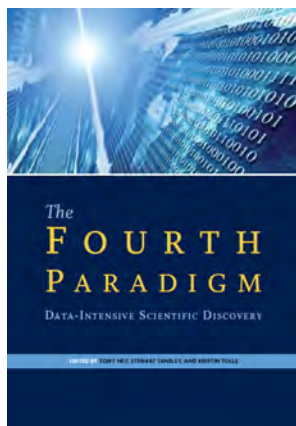
1

本日の講義

1. はじめに：なぜデータに注目するのか？
2. データベースと知識ベース
3. 生命科学分野におけるデータベースの事始め
4. 異なるデータを接続する
5. 異なるデータベースを接続する
6. おわりに

「計算生命科学の基礎Ⅴ」

2



Microsoft Research (2009)

はじめに なぜデータに注目するのか？

科学の発展（知識獲得の欲求）

観測から得られる経験則

経験則からの理論構築

理論からシミュレーションへ

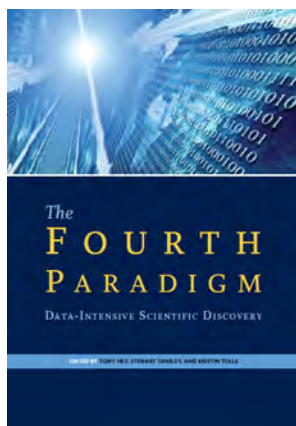
データを大量に
産出するようになる

観測、経験則、理論、シミュレーション
を接続する（理解する）



データサイエンス：データの関係を知る

「計算生命科学の基礎Ⅴ」



Microsoft Research (2009)

はじめに： なぜデータに注目するのか？

科学の発展（知識獲得の欲求）

生命科学のデータサイエンス：

生命科学データの関係を知る

計算能力の向上

bioinformatics（データ）と

computational biology（シミュレーション）の誕生

俯瞰的理解を求めることが
できるようになる。

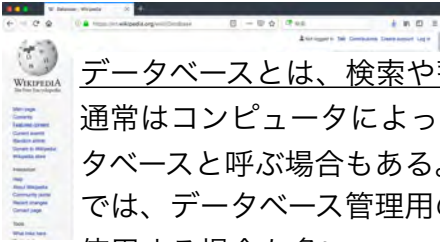


データの蓄積と知識の蓄積が必要

「計算生命科学の基礎Ⅴ」



データベースと知識ベース

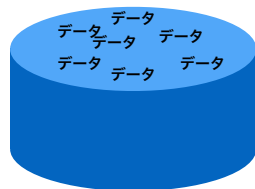


データベースとは、検索や蓄積が容易にできるよう整理された情報の集まり。
通常はコンピュータによって実現されたものを指すが、紙の住所録などをデータベースと呼ぶ場合もある。コンピュータを使用したデータベース・システムでは、データベース管理用のソフトウェアであるデータベース管理システムを使用する場合も多い。

<https://ja.wikipedia.org/wiki/データベース>

A database is an organized collection of data, generally stored and accessed electronically from a computer system. Where databases are more complex they are often developed using formal design and modeling techniques.

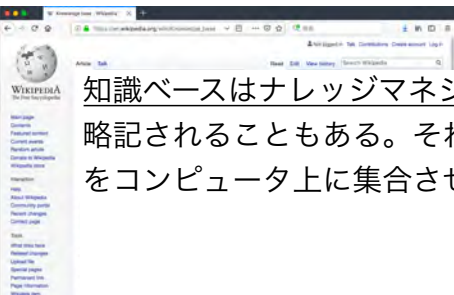
<https://en.wikipedia.org/wiki/Database>



データが存在する
データを探することができる



データベースと知識ベース

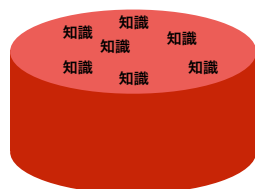


知識ベースはナレッジマネジメントのための特殊なデータベースであり、KBと略記されることもある。それは知識の検索を可能とし、知識を組織化し、知識をコンピュータ上に集合させたものである。

<https://ja.wikipedia.org/wiki/知識ベース>

A knowledge base (KB) is a technology used to store complex structured and unstructured information used by a computer system. The initial use of the term was in connection with expert systems which were the first knowledge-based systems.

https://en.wikipedia.org/wiki/Knowledge_base



知識が存在する
知識を探することができる

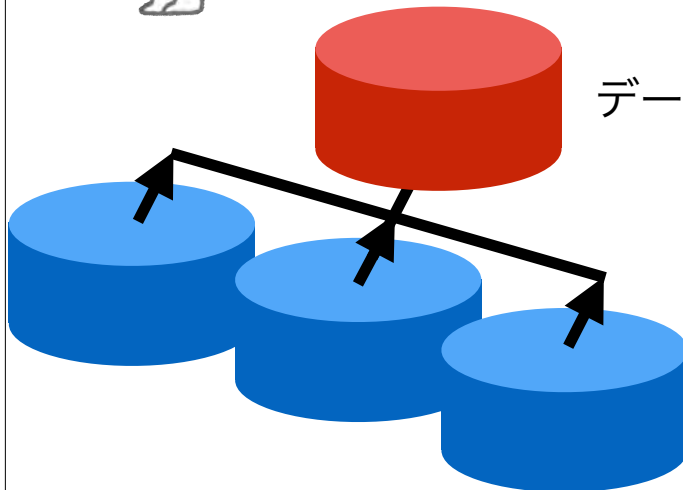
データベースと知識ベース



生命科学の
問い合わせ



<https://en.wikipedia.org/wiki/LCARS>



データ間のルールが存在する

データが存在する

「計算生命科学の基礎Ⅴ」

この仕組みを実現することをめざす

7

生命科学分野における データベースの事始め

時代背景

生命科学分野の初期のデータベース

- 1971年：タンパク質立体構造データベースPDBがアメリカ合衆国で設立
- 1972年：タンパク質アミノ酸配列データベース(NBRF/PIR)がWashington DCで設立
- 1979年：核酸配列データベースがアメリカ合衆国Los Alamosで設立
- 1980年：核酸配列データベースが西ドイツ・EMBLで設立
- 1984年：核酸配列データベースが日本・遺伝学研究所で設立
- 1989年：ゲノムデータベース管理システムの登場(AceDB)
- 1995年：Haemophilus influenzaeのゲノム全塩基配列決定
- 1997年：PROSITEデータベースの設立
- 1996年：パスイデータベースKEGGの設立
- 2000年：遺伝子オントロジーコンソーシアムの設立

- 1943年：Colossus稼働
- 1951年：一次構造発表
- 1953年：DNA二重らせん構造発表
- 1957年：三次構造発表
- 1966年：64種の遺伝暗号解読完了
- 1970年：関係データベースの提案

- 1972年：PAM250発表
- 1973年：UNIX OS誕生

- 1986年：ヒトゲノムプロジェクトの提案
- 1992年：World Wide Web の登場

- 1995年：マイクロアレイ法の開発

- 2000年：ツークハイブリッド法による
出芽酵母全遺伝子間相互作用の解析

生命科学データベースは古くから存在する

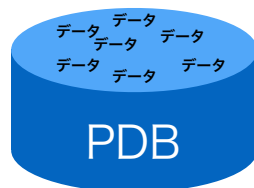
データベースにもとづく研究活動が生まれる

「計算生命科学の基礎Ⅴ」

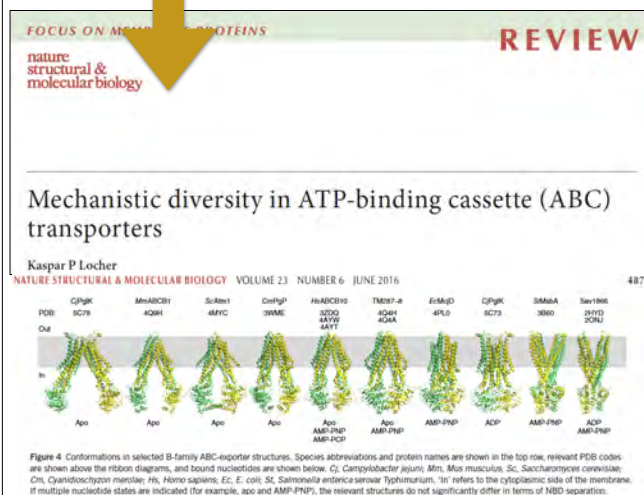
8

異なるデータを接続する (1)

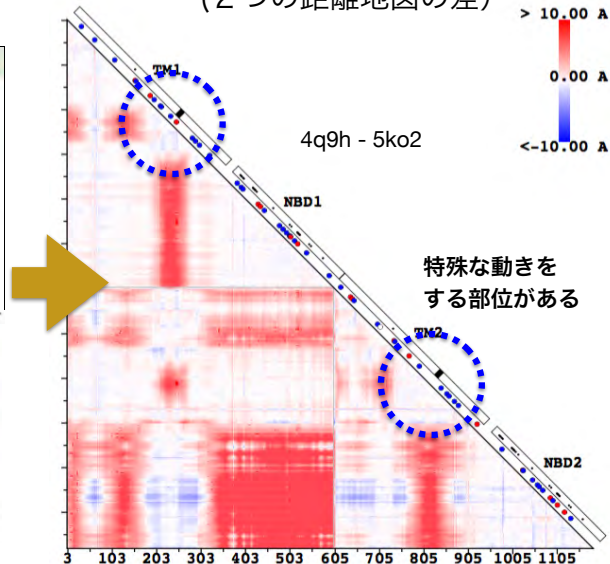
タンパク質の動きが見える



様々な状態の
立体構造情報



differential map
(2つの距離地図の差)

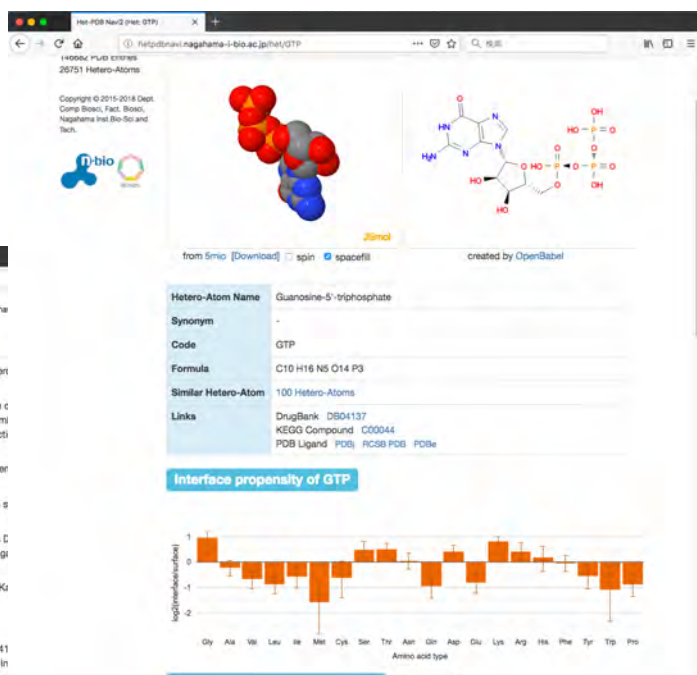
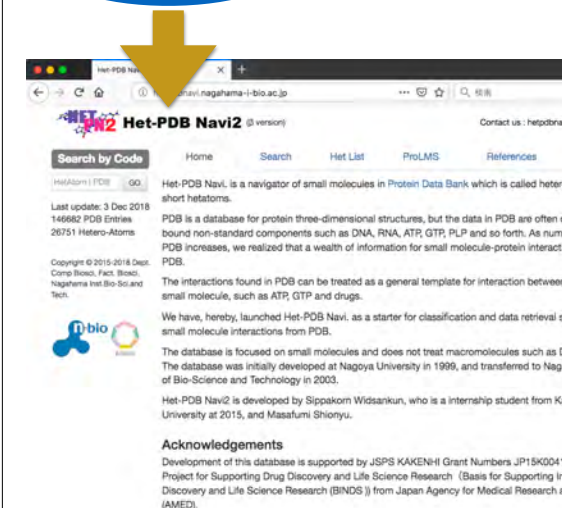
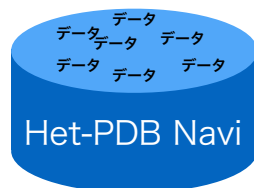


「計算生命科学の基礎V」

9

異なるデータを接続する (1)

タンパク質に結合する低分子を集める

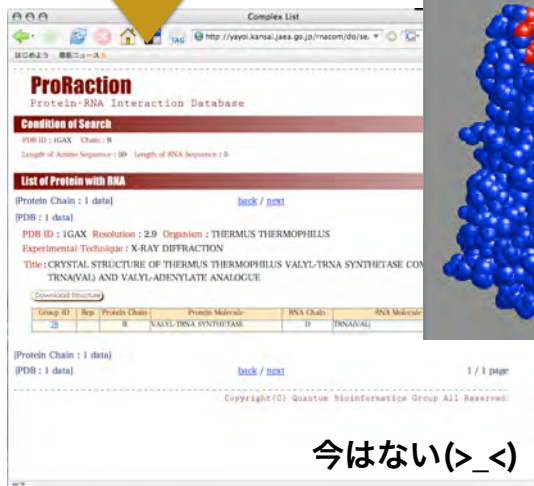
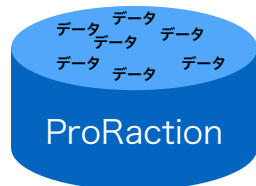


GTPはどのようなアミノ酸残基で認識されているか？

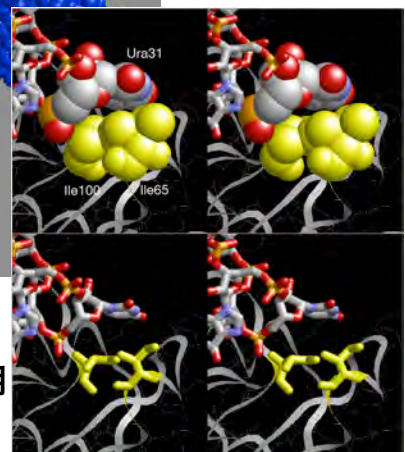
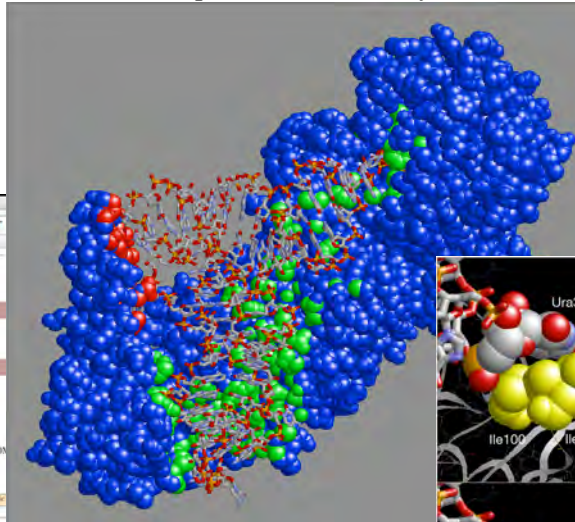
「計算生命科学の基礎V」

異なるデータを接続する (1)

RNAを結合するタンパク質を集める



Thermus thermophilus Val-tRNA synthetase



今はない(>_<)

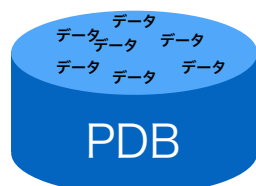
新規の相互作用
が見つかる

「計算生命科学の基礎V」

11

異なるデータを接続する (2)

タンパク質立体構造を推定する



タンパク質の機能発現原理をタンパク質立体構造に基づいて理解するにあたり、立体構造情報が必要だが、その情報を計測することが非常に困難である。

Proc. Natl. Acad. Sci. USA
Vol. 75, No. 1, pp. 184-188, January 1978
Biochemistry

Insulin-like growth factor: A model for tertiary structure accounting for immunoreactivity and receptor binding

(nonsuppressible insulin-like activity/molecular model building/evolution of insulin/amino acid sequence homology/quaternary structure)

T. L. BLUNDELL*, S. BEDARKAR*, E. RINDERKNECHT†, AND R. E. HUMBEL†

* Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom; and † Biochemisches Institut der Universität Zürich, CH-8028 Zürich, Switzerland

Communicated by Donald F. Steiner, October 25, 1977

ABSTRACT A model for the three-dimensional structure of insulin-like growth factor (IGF) is proposed based on the close sequence homology of IGF with insulin, the tertiary structure of which is known. The IGF molecule is postulated to have an insulin-like main chain conformation for residues equivalent to B6-B27 and A1-A21 and a hydrophobic core nearly identical to that of insulin. A short connecting peptide of twelve residues and an extension at the COOH-terminus are easily accommodated on the molecular surface. The surface involved in dimer formation in insulin is largely conserved, but the zinc-binding histidine and many residues involving hexamerization are very different from those of insulin and it is unlikely that IGF forms zinc hexamers. The model provides a ready explanation for the inability of IGF to bind antibodies to insulin and for its ability to bind insulin receptors with low affinity.

puter program of Isaacs. and other intramolecular distances were calculated using an IBM 360/65 computer. The model was then displayed on a computer graphics system (Digital Equipment Corp. graphics with a PDP11 computer) using programs written by D. Richardson, P. Pauling, and C. Chothia, and adjustments were made using the interactive facilities of the graphics system to optimize the intramolecular distances. Finally, the model was regeometricized using "modelfit" and stereo pairs of the model were generated in hard copy.

RESULTS AND DISCUSSION

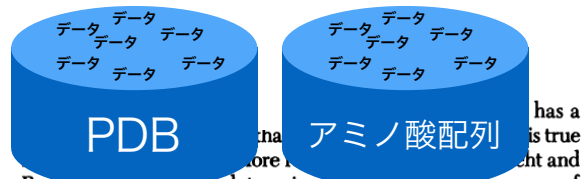
初めてのホモロジーモデリング

「計算生命科学の基礎V」

12

異なるデータを接続する (2)

タンパク質立体構造を推定する

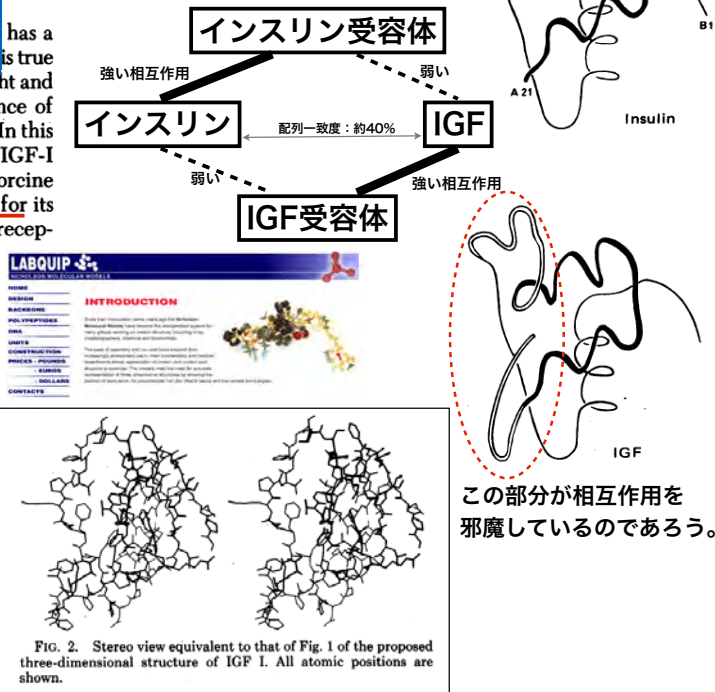


has a true structure. R. E. Hawley, have determined the complete sequence of IGF-I, which shows clearly homology with proinsulin. In this paper we present a model for the tertiary structure of IGF-I based on the known three-dimensional structure of porcine insulin (10, 11) and show how this model may account for its immunoreactivity and its affinity for insulin and IGF receptors.

MATERIALS AND METHODS

The model was built in several stages. The proposed structure was first constructed using Lapquip model parts at a scale of 1 cm = 1 Å. This was examined for unfavorable intramolecular contracts and readjusted before the coordinates of each atom were read off using a mechanical device. The approximate coordinates were then regularized using the "modelfit" computer program of Isaacs *et al* (12), and bond lengths and angles and other intramolecular distances were calculated using an IBM 360/65 computer. The model was then displayed on a computer graphics system (Digital Equipment Corp. graphics with a PDP11 computer) using programs written by D. Richardson, P. Pauling, and C. Chothia, and adjustments were made using the interactive facilities of the graphics system to optimize the intramolecular distances. Finally, the model was regeometricized using "modelfit" and stereo pairs of the model were generated in hard copy.

「計算生命科学の基礎V」



13

異なるデータを接続する (2)

MSH6に見られるガンの原因変異は?



MSH6アミノ酸配列

```

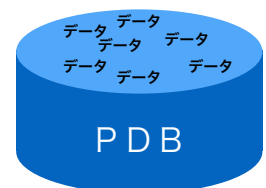
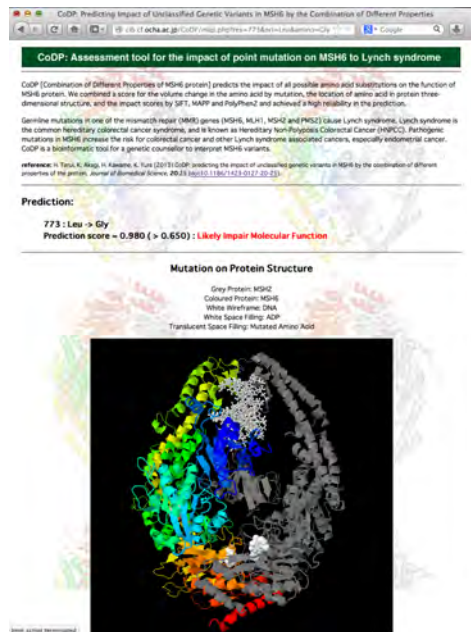
1  MSH6 1-1000
2  1  MSH6 1-1000
3  1  MSH6 1-1000
4  1  MSH6 1-1000
5  1  MSH6 1-1000
6  1  MSH6 1-1000
7  1  MSH6 1-1000
8  1  MSH6 1-1000
9  1  MSH6 1-1000
10 1  MSH6 1-1000
11 1  MSH6 1-1000
12 1  MSH6 1-1000
13 1  MSH6 1-1000
14 1  MSH6 1-1000
15 1  MSH6 1-1000
16 1  MSH6 1-1000
17 1  MSH6 1-1000
18 1  MSH6 1-1000
19 1  MSH6 1-1000
20 1  MSH6 1-1000
21 1  MSH6 1-1000
22 1  MSH6 1-1000
23 1  MSH6 1-1000
24 1  MSH6 1-1000
25 1  MSH6 1-1000
26 1  MSH6 1-1000
27 1  MSH6 1-1000
28 1  MSH6 1-1000
29 1  MSH6 1-1000
30 1  MSH6 1-1000
31 1  MSH6 1-1000
32 1  MSH6 1-1000
33 1  MSH6 1-1000
34 1  MSH6 1-1000
35 1  MSH6 1-1000
36 1  MSH6 1-1000
37 1  MSH6 1-1000
38 1  MSH6 1-1000
39 1  MSH6 1-1000
40 1  MSH6 1-1000
41 1  MSH6 1-1000
42 1  MSH6 1-1000
43 1  MSH6 1-1000
44 1  MSH6 1-1000
45 1  MSH6 1-1000
46 1  MSH6 1-1000
47 1  MSH6 1-1000
48 1  MSH6 1-1000
49 1  MSH6 1-1000
50 1  MSH6 1-1000
51 1  MSH6 1-1000
52 1  MSH6 1-1000
53 1  MSH6 1-1000
54 1  MSH6 1-1000
55 1  MSH6 1-1000
56 1  MSH6 1-1000
57 1  MSH6 1-1000
58 1  MSH6 1-1000
59 1  MSH6 1-1000
60 1  MSH6 1-1000
61 1  MSH6 1-1000
62 1  MSH6 1-1000
63 1  MSH6 1-1000
64 1  MSH6 1-1000
65 1  MSH6 1-1000
66 1  MSH6 1-1000
67 1  MSH6 1-1000
68 1  MSH6 1-1000
69 1  MSH6 1-1000
70 1  MSH6 1-1000
71 1  MSH6 1-1000
72 1  MSH6 1-1000
73 1  MSH6 1-1000
74 1  MSH6 1-1000
75 1  MSH6 1-1000
76 1  MSH6 1-1000
77 1  MSH6 1-1000
78 1  MSH6 1-1000
79 1  MSH6 1-1000
80 1  MSH6 1-1000
81 1  MSH6 1-1000
82 1  MSH6 1-1000
83 1  MSH6 1-1000
84 1  MSH6 1-1000
85 1  MSH6 1-1000
86 1  MSH6 1-1000
87 1  MSH6 1-1000
88 1  MSH6 1-1000
89 1  MSH6 1-1000
90 1  MSH6 1-1000
91 1  MSH6 1-1000
92 1  MSH6 1-1000
93 1  MSH6 1-1000
94 1  MSH6 1-1000
95 1  MSH6 1-1000
96 1  MSH6 1-1000
97 1  MSH6 1-1000
98 1  MSH6 1-1000
99 1  MSH6 1-1000
100 1  MSH6 1-1000
101 1  MSH6 1-1000
102 1  MSH6 1-1000
103 1  MSH6 1-1000
104 1  MSH6 1-1000
105 1  MSH6 1-1000
106 1  MSH6 1-1000
107 1  MSH6 1-1000
108 1  MSH6 1-1000
109 1  MSH6 1-1000
110 1  MSH6 1-1000
111 1  MSH6 1-1000
112 1  MSH6 1-1000
113 1  MSH6 1-1000
114 1  MSH6 1-1000
115 1  MSH6 1-1000
116 1  MSH6 1-1000
117 1  MSH6 1-1000
118 1  MSH6 1-1000
119 1  MSH6 1-1000
120 1  MSH6 1-1000
121 1  MSH6 1-1000
122 1  MSH6 1-1000
123 1  MSH6 1-1000
124 1  MSH6 1-1000
125 1  MSH6 1-1000
126 1  MSH6 1-1000
127 1  MSH6 1-1000
128 1  MSH6 1-1000
129 1  MSH6 1-1000
130 1  MSH6 1-1000
131 1  MSH6 1-1000
132 1  MSH6 1-1000
133 1  MSH6 1-1000
134 1  MSH6 1-1000
135 1  MSH6 1-1000
136 1  MSH6 1-1000
137 1  MSH6 1-1000
138 1  MSH6 1-1000
139 1  MSH6 1-1000
140 1  MSH6 1-1000
141 1  MSH6 1-1000
142 1  MSH6 1-1000
143 1  MSH6 1-1000
144 1  MSH6 1-1000
145 1  MSH6 1-1000
146 1  MSH6 1-1000
147 1  MSH6 1-1000
148 1  MSH6 1-1000
149 1  MSH6 1-1000
150 1  MSH6 1-1000
151 1  MSH6 1-1000
152 1  MSH6 1-1000
153 1  MSH6 1-1000
154 1  MSH6 1-1000
155 1  MSH6 1-1000
156 1  MSH6 1-1000
157 1  MSH6 1-1000
158 1  MSH6 1-1000
159 1  MSH6 1-1000
160 1  MSH6 1-1000
161 1  MSH6 1-1000
162 1  MSH6 1-1000
163 1  MSH6 1-1000
164 1  MSH6 1-1000
165 1  MSH6 1-1000
166 1  MSH6 1-1000
167 1  MSH6 1-1000
168 1  MSH6 1-1000
169 1  MSH6 1-1000
170 1  MSH6 1-1000
171 1  MSH6 1-1000
172 1  MSH6 1-1000
173 1  MSH6 1-1000
174 1  MSH6 1-1000
175 1  MSH6 1-1000
176 1  MSH6 1-1000
177 1  MSH6 1-1000
178 1  MSH6 1-1000
179 1  MSH6 1-1000
180 1  MSH6 1-1000
181 1  MSH6 1-1000
182 1  MSH6 1-1000
183 1  MSH6 1-1000
184 1  MSH6 1-1000
185 1  MSH6 1-1000
186 1  MSH6 1-1000
187 1  MSH6 1-1000
188 1  MSH6 1-1000
189 1  MSH6 1-1000
190 1  MSH6 1-1000
191 1  MSH6 1-1000
192 1  MSH6 1-1000
193 1  MSH6 1-1000
194 1  MSH6 1-1000
195 1  MSH6 1-1000
196 1  MSH6 1-1000
197 1  MSH6 1-1000
198 1  MSH6 1-1000
199 1  MSH6 1-1000
200 1  MSH6 1-1000
201 1  MSH6 1-1000
202 1  MSH6 1-1000
203 1  MSH6 1-1000
204 1  MSH6 1-1000
205 1  MSH6 1-1000
206 1  MSH6 1-1000
207 1  MSH6 1-1000
208 1  MSH6 1-1000
209 1  MSH6 1-1000
210 1  MSH6 1-1000
211 1  MSH6 1-1000
212 1  MSH6 1-1000
213 1  MSH6 1-1000
214 1  MSH6 1-1000
215 1  MSH6 1-1000
216 1  MSH6 1-1000
217 1  MSH6 1-1000
218 1  MSH6 1-1000
219 1  MSH6 1-1000
220 1  MSH6 1-1000
221 1  MSH6 1-1000
222 1  MSH6 1-1000
223 1  MSH6 1-1000
224 1  MSH6 1-1000
225 1  MSH6 1-1000
226 1  MSH6 1-1000
227 1  MSH6 1-1000
228 1  MSH6 1-1000
229 1  MSH6 1-1000
230 1  MSH6 1-1000
231 1  MSH6 1-1000
232 1  MSH6 1-1000
233 1  MSH6 1-1000
234 1  MSH6 1-1000
235 1  MSH6 1-1000
236 1  MSH6 1-1000
237 1  MSH6 1-1000
238 1  MSH6 1-1000
239 1  MSH6 1-1000
240 1  MSH6 1-1000
241 1  MSH6 1-1000
242 1  MSH6 1-1000
243 1  MSH6 1-1000
244 1  MSH6 1-1000
245 1  MSH6 1-1000
246 1  MSH6 1-1000
247 1  MSH6 1-1000
248 1  MSH6 1-1000
249 1  MSH6 1-1000
250 1  MSH6 1-1000
251 1  MSH6 1-1000
252 1  MSH6 1-1000
253 1  MSH6 1-1000
254 1  MSH6 1-1000
255 1  MSH6 1-1000
256 1  MSH6 1-1000
257 1  MSH6 1-1000
258 1  MSH6 1-1000
259 1  MSH6 1-1000
260 1  MSH6 1-1000
261 1  MSH6 1-1000
262 1  MSH6 1-1000
263 1  MSH6 1-1000
264 1  MSH6 1-1000
265 1  MSH6 1-1000
266 1  MSH6 1-1000
267 1  MSH6 1-1000
268 1  MSH6 1-1000
269 1  MSH6 1-1000
270 1  MSH6 1-1000
271 1  MSH6 1-1000
272 1  MSH6 1-1000
273 1  MSH6 1-1000
274 1  MSH6 1-1000
275 1  MSH6 1-1000
276 1  MSH6 1-1000
277 1  MSH6 1-1000
278 1  MSH6 1-1000
279 1  MSH6 1-1000
280 1  MSH6 1-1000
281 1  MSH6 1-1000
282 1  MSH6 1-1000
283 1  MSH6 1-1000
284 1  MSH6 1-1000
285 1  MSH6 1-1000
286 1  MSH6 1-1000
287 1  MSH6 1-1000
288 1  MSH6 1-1000
289 1  MSH6 1-1000
290 1  MSH6 1-1000
291 1  MSH6 1-1000
292 1  MSH6 1-1000
293 1  MSH6 1-1000
294 1  MSH6 1-1000
295 1  MSH6 1-1000
296 1  MSH6 1-1000
297 1  MSH6 1-1000
298 1  MSH6 1-1000
299 1  MSH6 1-1000
300 1  MSH6 1-1000
301 1  MSH6 1-1000
302 1  MSH6 1-1000
303 1  MSH6 1-1000
304 1  MSH6 1-1000
305 1  MSH6 1-1000
306 1  MSH6 1-1000
307 1  MSH6 1-1000
308 1  MSH6 1-1000
309 1  MSH6 1-1000
310 1  MSH6 1-1000
311 1  MSH6 1-1000
312 1  MSH6 1-1000
313 1  MSH6 1-1000
314 1  MSH6 1-1000
315 1  MSH6 1-1000
316 1  MSH6 1-1000
317 1  MSH6 1-1000
318 1  MSH6 1-1000
319 1  MSH6 1-1000
320 1  MSH6 1-1000
321 1  MSH6 1-1000
322 1  MSH6 1-1000
323 1  MSH6 1-1000
324 1  MSH6 1-1000
325 1  MSH6 1-1000
326 1  MSH6 1-1000
327 1  MSH6 1-1000
328 1  MSH6 1-1000
329 1  MSH6 1-1000
330 1  MSH6 1-1000
331 1  MSH6 1-1000
332 1  MSH6 1-1000
333 1  MSH6 1-1000
334 1  MSH6 1-1000
335 1  MSH6 1-1000
336 1  MSH6 1-1000
337 1  MSH6 1-1000
338 1  MSH6 1-1000
339 1  MSH6 1-1000
340 1  MSH6 1-1000
341 1  MSH6 1-1000
342 1  MSH6 1-1000
343 1  MSH6 1-1000
344 1  MSH6 1-1000
345 1  MSH6 1-1000
346 1  MSH6 1-1000
347 1  MSH6 1-1000
348 1  MSH6 1-1000
349 1  MSH6 1-1000
350 1  MSH6 1-1000
351 1  MSH6 1-1000
352 1  MSH6 1-1000
353 1  MSH6 1-1000
354 1  MSH6 1-1000
355 1  MSH6 1-1000
356 1  MSH6 1-1000
357 1  MSH6 1-1000
358 1  MSH6 1-1000
359 1  MSH6 1-1000
360 1  MSH6 1-1000
361 1  MSH6 1-1000
362 1  MSH6 1-1000
363 1  MSH6 1-1000
364 1  MSH6 1-1000
365 1  MSH6 1-1000
366 1  MSH6 1-1000
367 1  MSH6 1-1000
368 1  MSH6 1-1000
369 1  MSH6 1-1000
370 1  MSH6 1-1000
371 1  MSH6 1-1000
372 1  MSH6 1-1000
373 1  MSH6 1-1000
374 1  MSH6 1-1000
375 1  MSH6 1-1000
376 1  MSH6 1-1000
377 1  MSH6 1-1000
378 1  MSH6 1-1000
379 1  MSH6 1-1000
380 1  MSH6 1-1000
381 1  MSH6 1-1000
382 1  MSH6 1-1000
383 1  MSH6 1-1000
384 1  MSH6 1-1000
385 1  MSH6 1-1000
386 1  MSH6 1-1000
387 1  MSH6 1-1000
388 1  MSH6 1-1000
389 1  MSH6 1-1000
390 1  MSH6 1-1000
391 1  MSH6 1-1000
392 1  MSH6 1-1000
393 1  MSH6 1-1000
394 1  MSH6 1-1000
395 1  MSH6 1-1000
396 1  MSH6 1-1000
397 1  MSH6 1-1000
398 1  MSH6 1-1000
399 1  MSH6 1-1000
400 1  MSH6 1-1000
401 1  MSH6 1-1000
402 1  MSH6 1-1000
403 1  MSH6 1-1000
404 1  MSH6 1-1000
405 1  MSH6 1-1000
406 1  MSH6 1-1000
407 1  MSH6 1-1000
408 1  MSH6 1-1000
409 1  MSH6 1-1000
410 1  MSH6 1-1000
411 1  MSH6 1-1000
412 1  MSH6 1-1000
413 1  MSH6 1-1000
414 1  MSH6 1-1000
415 1  MSH6 1-1000
416 1  MSH6 1-1000
417 1  MSH6 1-1000
418 1  MSH6 1-1000
419 1  MSH6 1-1000
420 1  MSH6 1-1000
421 1  MSH6 1-1000
422 1  MSH6 1-1000
423 1  MSH6 1-1000
424 1  MSH6 1-1000
425 1  MSH6 1-1000
426 1  MSH6 1-1000
427 1  MSH6 1-1000
428 1  MSH6 1-1000
429 1  MSH6 1-1000
430 1  MSH6 1-1000
431 1  MSH6 1-1000
432 1  MSH6 1-1000
433 1  MSH6 1-1000
434 1  MSH6 1-1000
435 1  MSH6 1-1000
436 1  MSH6 1-1000
437 1  MSH6 1-1000
438 1  MSH6 1-1000
439 1  MSH6 1-1000
440 1  MSH6 1-1000
441 1  MSH6 1-1000
442 1  MSH6 1-1000
443 1  MSH6 1-1000
444 1  MSH6 1-1000
445 1  MSH6 1-1000
446 1  MSH6 1-1000
447 1  MSH6 1-1000
448 1  MSH6 1-1000
449 1  MSH6 1-1000
450 1  MSH6 1-1000
451 1  MSH6 1-1000
452 1  MSH6 1-1000
453 1  MSH6 1-1000
454 1  MSH6 1-1000
455 1  MSH6 1-1000
456 1  MSH6 1-1000
457 1  MSH6 1-1000
458 1  MSH6 1-1000
459 1  MSH6 1-1000
460 1  MSH6 1-1000
461 1  MSH6 1-1000
462 1  MSH6 1-1000
463 1  MSH6 1-1000
464 1  MSH6 1-1000
465 1  MSH6 1-1000
466 1  MSH6 1-1000
467 1  MSH6 1-1000
468 1  MSH6 1-1000
469 1  MSH6 1-1000
470 1  MSH6 1-1000
471 1  MSH6 1-1000
472 1  MSH6 1-1000
473 1  MSH6 1-1000
474 1  MSH6 1-1000
475 1  MSH6 1-1000
476 1  MSH6 1-1000
477 1  MSH6 1-1000
478 1  MSH6 1-1000
479 1  MSH6 1-1000
480 1  MSH6 1-1000
481 1  MSH6 1-1000
482 1  MSH6 1-1000
483 1  MSH6 1-1000
484 1  MSH6 1-1000
485 1  MSH6 1-1000
486 1  MSH6 1-1000
487 1  MSH6 1-1000
488 1  MSH6 1-1000
489 1  MSH6 1-1000
490 1  MSH6 1-1000
491 1  MSH6 1-1000
492 1  MSH6 1-1000
493 1  MSH6 1-1000
494 1  MSH6 1-1000
495 1  MSH6 1-1000
496 1  MSH6 1-1000
497 1  MSH6 1-1000
498 1  MSH6 1-1000
499 1  MSH6 1-1000
500 1  MSH6 1-1000
501 1  MSH6 1-1000
502 1  MSH6 1-1000
503 1  MSH6 1-1000
504 1  MSH6 1-1000
505 1  MSH6 1-1000
506 1  MSH6 1-1000
507 1  MSH6 1-1000
508 1  MSH6 1-1000
509 1  MSH6 1-1000
510 1  MSH6 1-1000
511 1  MSH6 1-1000
512 1  MSH6 1-1000
513 1  MSH6 1-1000
514 1  MSH6 1-1000
515 1  MSH6 1-1000
516 1  MSH6 1-1000
517 1  MSH6 1-1000
518 1  MSH6 1-1000
519 1  MSH6 1-1000
520 1  MSH6 1-1000
521 1  MSH6 1-1000
522 1  MSH6 1-1000
523 1  MSH6 1-1000
524 1  MSH6 1-1000
525 1  MSH6 1-1000
526 1  MSH6 1-1000
527 1  MSH6 1-1000
528 1  MSH6 1-1000
529 1  MSH6 1-1000
530 1  MSH6 1-1000
531 1  MSH6 1-1000
532 1  MSH6 1-1000
533 1  MSH6 1-1000
534 1  MSH6 1-1000
535 1  MSH6 1-1000
536 1  MSH6 1-1000
537 1  MSH6 1-1000
538 1  MSH6 1-1000
539 1  MSH6 1-1000
540 1  MSH6 1-1000
541 1  MSH6 1-1000
542 1  MSH6 1-1000
543 1  MSH6 1-1000
544 1  MSH6 1-1000
545 1  MSH6 1-1000
546 1  MSH6 1-1000
547 1  MSH6 1-1000
548 1  MSH6 1-1000
549 1  MSH6 1-1000
550 1  MSH6 1-1000
551 1  MSH6 1-1000
552 1  MSH6 1-1000
553 1  MSH6 1-1000
554 1  MSH6 1-1000
555 1  MSH6 1-1000
556 1  MSH6 1-1000
557 1  MSH6 1-1000
558 1  MSH6 1-1000
559 1  MSH6 1-1000
560 1  MSH6 1-1000
561 1  MSH6 1-1000
562 1  MSH6 1-1000
563 1  MSH6 1-1000
564 1  MSH6 1-1000
565 1  MSH6 1-1000
566 1  MSH6 1-1000
567 1  MSH6 1-1000
568 1  MSH6 1-1000
569 1  MSH6 1-1000
570 1  MSH6 1-1000
571 1  MSH6 1-1000
572 1  MSH6 1-1000
573 1  MSH6 1-1000
574 1  MSH6 1-1000
575 1  MSH6 1-1000
576 1  MSH6 1-1000
577 1  MSH6 1-1000
578 1  MSH6 1-1000
579 1  MSH6 1-1000
580 1  MSH6 1-1000
581 1  MSH6 1-1000
582 1  MSH6 1-1000
583 1  MSH6 1-1000
584 1  MSH6 1-1000
585 1  MSH6 1-1000
586 1  MSH6 1-1000
587 1  MSH6 1-1000
588 1  MSH6 1-1000
589 1  MSH6 1-1000
590 1  MSH6 1-1000
591 1  MSH6 1-1000
592 1  MSH6 1-1000
593 1  MSH6 1-1000
594 1  MSH6 1-1000
595 1  MSH6 1-1000
596 1  MSH6 1-1000
597 1  MSH6 1-1000
598 1  MSH6 1-1000
599 1  MSH6 1-1000
600 1  MSH6 1-1000
601 1  MSH6 1-1000
602 1  MSH6 1-1000
603 1  MSH6 1-1000
604 1  MSH6 1-1000
605 1  MSH6 1-1000
606 1  MSH6 1-1000
607 1  MSH6 1-1000
608 1  MSH6 1-1000
609 1  MSH6 1-1000
610 1  MSH6 1-1000
611 1  MSH6 1-1000
612 1  MSH6 1-1000
613 1  MSH6 1-1000
614 1  MSH6 1-1000
615 1  MSH6 1-1000
616 1  MSH6 1-1000
617 1  MSH6 1-1000
618 1  MSH6 1-1000
619 1  MSH6 1-1000
620 1  MSH6 1-1000
621 1  MSH6 1-1000
622 1  MSH6 1-1000
623 1  MSH6 1-1000
624 1  MSH6 1-1000
625 1  MSH6 1-1000
626 1  MSH6 1-1000
627 1  MSH6 1-1000
628 1  MSH6 1-1000
629 1  MSH6 1-1000
630 1  MSH6 1-1000
631 1  MSH6 1-1000
632 1  MSH6 1-1000
633 1  MSH6 1-1000
634 1  MSH6 1-1000
635 1  MSH6 1-1000
636 1  MSH6 1-1000
637 1  MSH6 1-1000
638 1  MSH6 1-1000
639 1  MSH6 1-1000
640 1  MSH6 1-1000
641 1  MSH6 1-1000
642 1  MSH6 1-1000
643 1  MSH6 1-1000
644 1  MSH6 1-1000
645 1  MSH6 1-1000
646 1  MSH6 1-1000
647 1  MSH6 1-1000
648 1  MSH6 1-1000
649 1  MSH6 1-1000
650 1  MSH6 1-1000
651 1  MSH6 1-1000
652 1  MSH6 1-1000
653 1  MSH6 1-1000
654 1  MSH6 1-1
```


異なるデータを接続する (2)

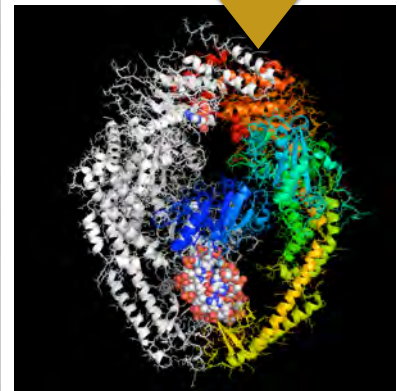
MSH6に見られるガンの原因変異は？



<http://cib.cf.ocha.ac.jp/CoDP/>



MSH6立体構造



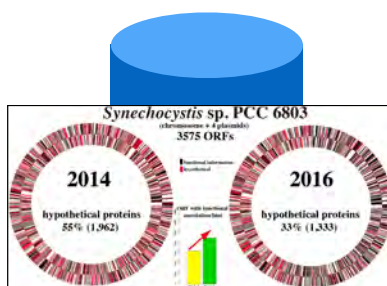
「計算生命科学の基礎V」

変異部位の特徴から影響を予測する

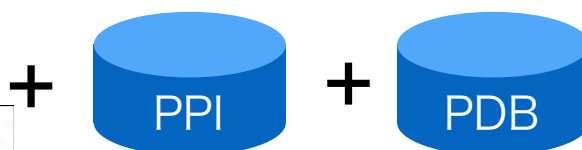
15

異なるデータを接続する (2)

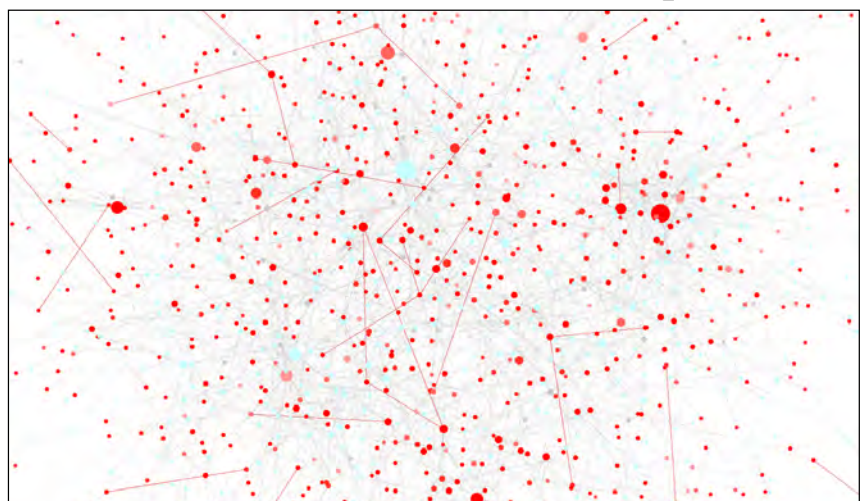
原子分解能で生命のしくみを理解する



アミノ酸配列と機能



巨大な「知」が描けた



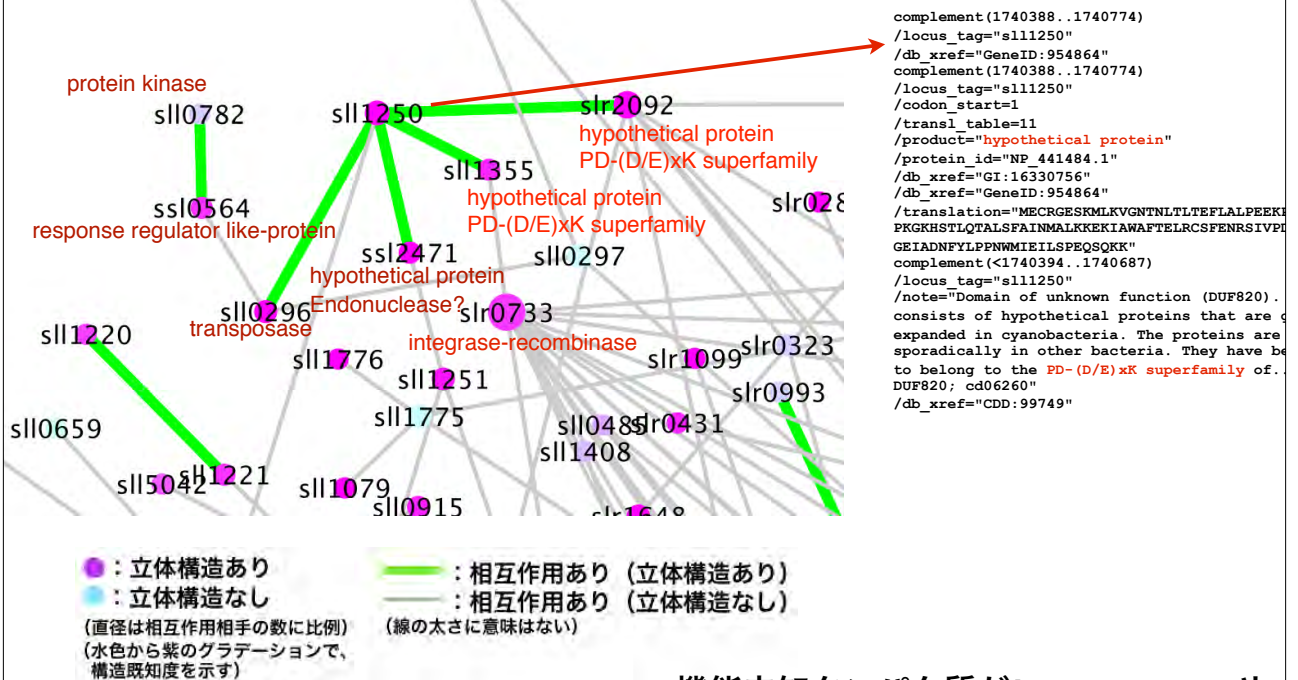
「計算生命科学の基礎V」

16



異なるデータを接続する (2)

原子分解能で生命のしくみを理解する



機能未知タンパク質がtransposaseや
integraseとつながっている。

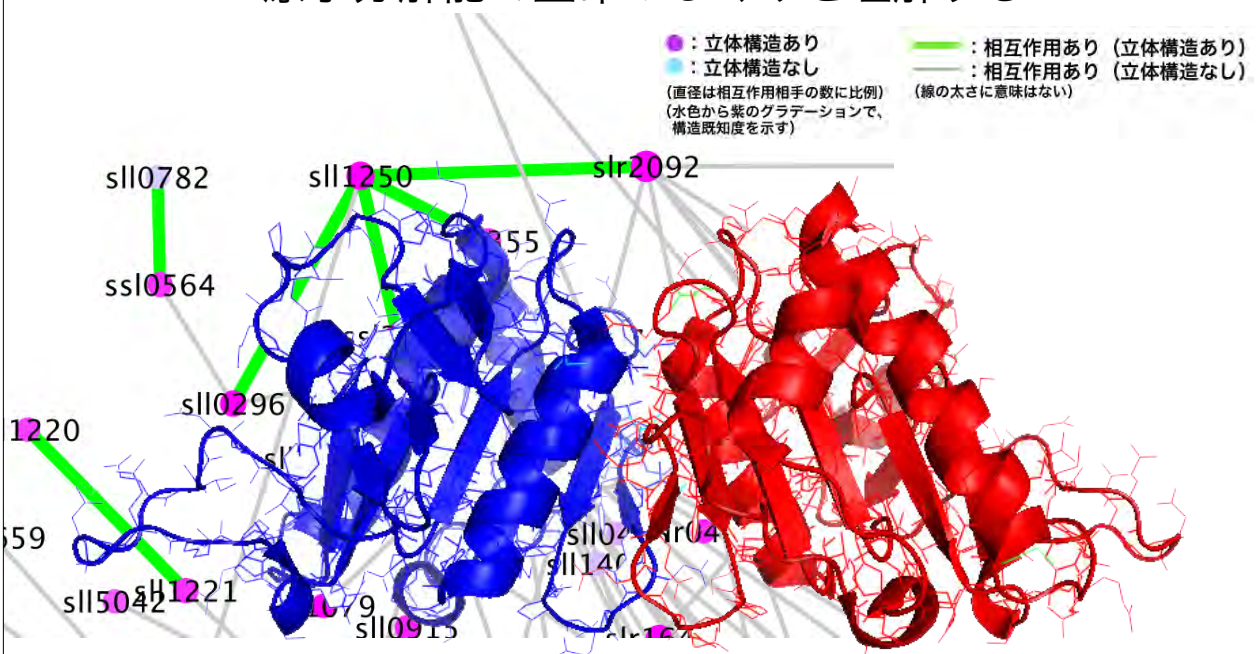
「計算生命科学の基礎V」

17



異なるデータを接続する (2)

原子分解能で生命のしくみを理解する



新たなタンパク質群がDNAの変化に
関与しているかもしれない

「計算生命科学の基礎V」

18

ここまでのデータベース

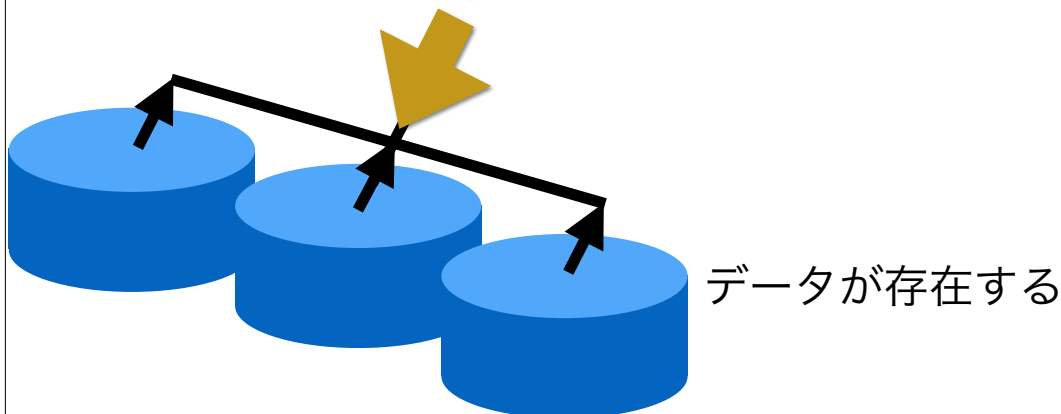
生命科学の問い合わせ



データ間のルールを
自ら探せ。



<https://en.wikipedia.org/wiki/LCARS>



「計算生命科学の基礎V」

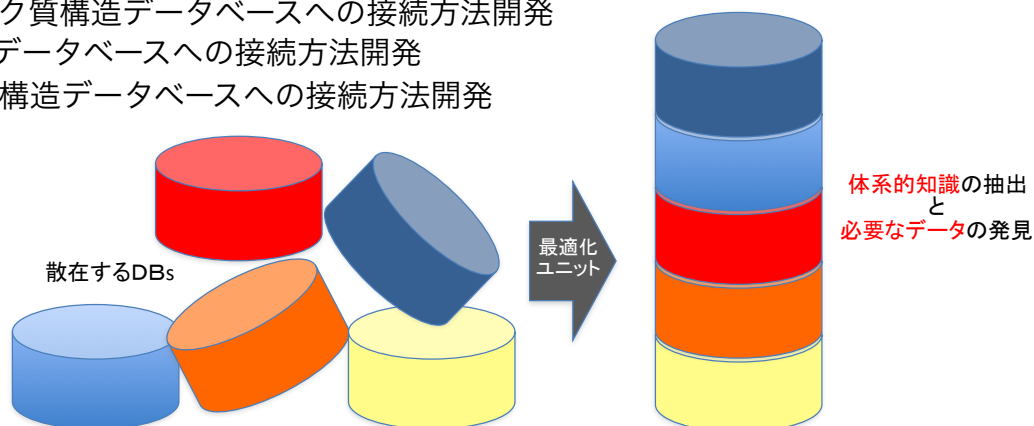
19

異なるデータベースを接続する

「機能最適化最適化ユニット」のミッション

- ・新しいデータベースとツールの開発
 - クライオ電子顕微鏡測定結果集積データベースの開発
 - 分子動力学シミュレーション結果集積データベースの開発
 - 構造解析ターゲティングデータベースの開発
 - ゲノム変異のタンパク質アノテーションツールの開発
- ・分散データベースからの知識取得方法 (VaProS)の開発
 - 知識取得方法の開発
 - タンパク質構造データベースへの接続方法開発
 - ゲノムデータベースへの接続方法開発
 - 超分子構造データベースへの接続方法開発

<http://www.binds.jp/>

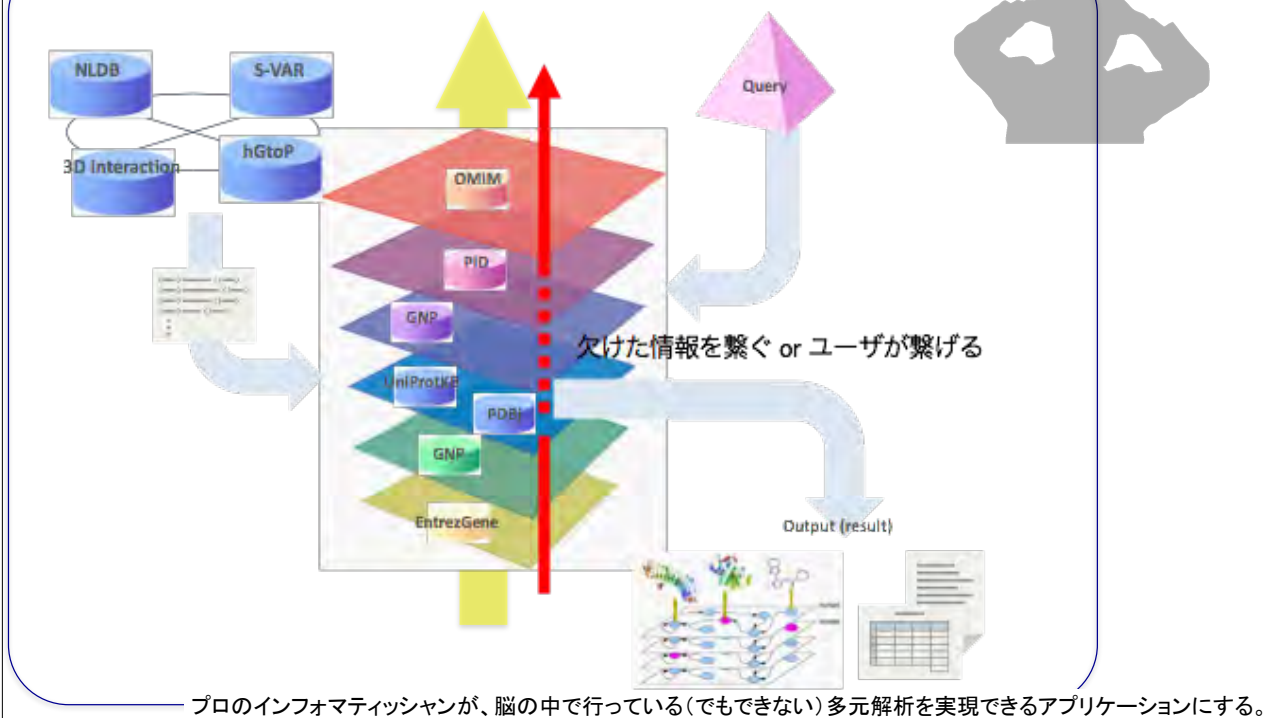


「計算生命科学の基礎V」

20

異なるデータベースを接続する

データの取得方法、データの読み方、データの解析方法などの困難から解放され、知識(データの関係)を発見するアプリケーションに育て上げる。



「計算生命科学の基礎Ⅴ」

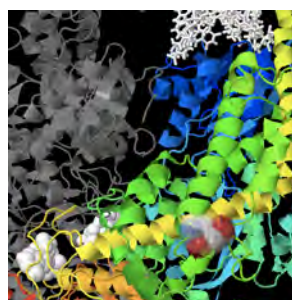
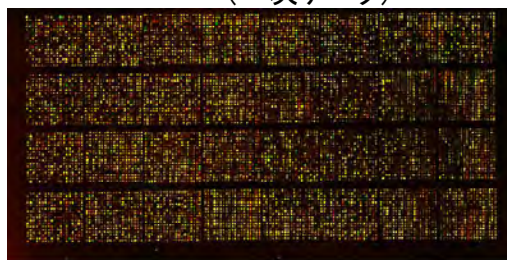
21

異なるデータベースを接続する

病気原因遺伝子を探そう

癌患者と健康者の遺伝子発現データ
(一次データ)

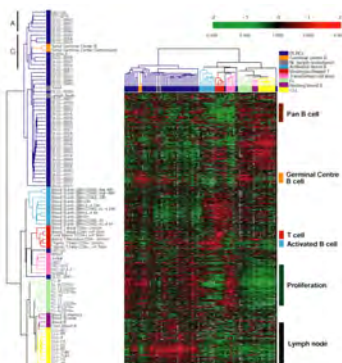
タンパク質の相互作用データ(二次データ)



遺伝子発現データ(二次データ)

タンパク質立体構造データ
(一次データ)

遺伝子変異のデータ(一次データ)



データの取得方法
データの読み方
データの解析方法

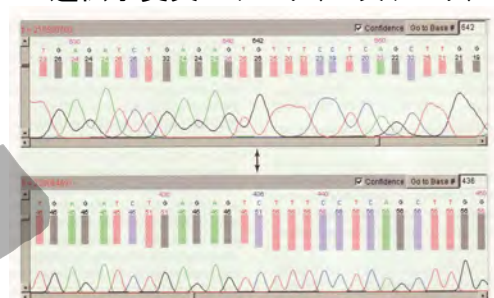
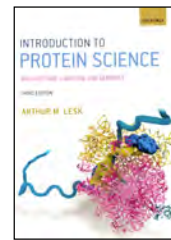


FIGURE 7.1 A large fraction of the SNPs in dbSNP were discovered from traces

「計算生命科学の基礎Ⅴ」

22

異なるデータベースを接続する 統合データベースの構築 vs データベースの統合



統合データベースの構築には、タンパク質が「アグutting」してしまうのと同じような危険がある。このデータベースの強みを失うことになるかもしれない。

データベースを統合するのは、技術的な困難があるが、データ統合のもう一つの手段であろう。興味深いアプローチだが、説かなければいけない、理論的・技術的問題が山積である。

Databases and information retrieval 161

between the surfaces, interiors and interfaces of proteins and protein complexes.

The eminent physical chemist E.A. Molwyn-Hughes wrote long ago that: 'The complete physical chemist blows his own apparatus and solves his own equations.' What would be the modern equivalent?

Perhaps: The complete protein scientist grows his or her own cells, sequences genes from these cells, isolates proteins and solves their structures, and writes his or her own computer programs to analyse how they relate to the totality of biological knowledge. A tall order!

Databases and information retrieval

Databases are central and crucial to contemporary biomedical research. The goals of the information enterprise are to archive, curate, and annotate data; to make the results available; and to impose a structure on the data that facilitates information retrieval. (A database without effective modes of access is merely a data graveyard.)

Information retrieval is easy if you know precisely what you are looking for: for instance, 'What is the amino acid sequence of adenylate kinase from *Micrococcus leuteus*?' The challenge is to allow the greatest generality in framing questions, even if the answers require access to combinations of different types of data, which may appear in different databases. These types of data might include protein sequences, muta-

There are dangers in database aggregation—analagous, to some extent, to the dangers of protein aggregation. One threat is loss of their individual strengths and the focused expertise of the curators of specialized data archiving projects. Of course it is possible to hire a cabinet of experts on different topics and set them to work down the hall from one another. In any event, the alternative to database consolidation is to devise standards for interoperability, so that a program could parse queries, farm out components to different databases, and collect and organize the results. That approach is an intriguing topic of theoretical research, but not yet a practical tool.

Indeed, at the moment, the trends towards consolidation appear much stronger than the centrif-

「計算生命科学の基礎Ⅴ」 VaProSが目指しているところ

23

異なるデータベースを接続する このような試みは数多くなされている

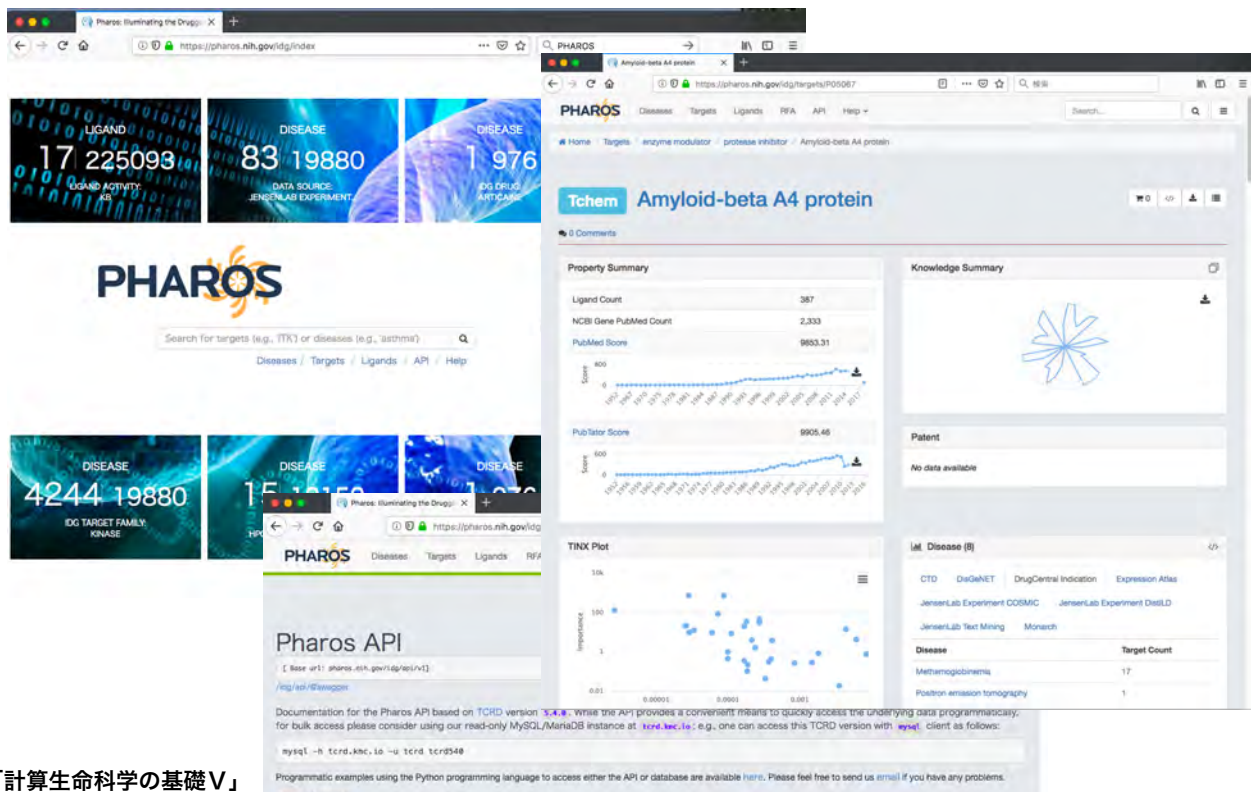
The image displays three web interfaces related to bioinformatics data integration. On the left is the DisGeNET website, which is a platform for gene-disease associations. In the center is the EMBL-EBI Services page, showing various tools and data resources available from the European Bioinformatics Institute. On the right is the Taverna Workbench Bioinformatics 2.5 download page, which provides information about the software and its system requirements.

「計算生命科学の基礎Ⅴ」

24

異なるデータベースを接続する

このような試みは数多くなされている



「計算生命科学の基礎V」

25

異なるデータベースを接続する



「計算生命科学の基礎V」

26

異なるデータベースを接続する



VaProS | Statistics

HOME ABOUT STATISTICS MANUAL TUTORIAL FAQ CONTACT



STATISTICS

Update time : Nov 02, 2017

Data Type	DB	Human	Mouse	Rat
Gene	EntrezGene	105,469	110,607	68,338
Protein	UniProtKB	524,973	310,123	113,005
	Swiss-Prot	1,388,713	560,341	321,312
Interactions	BioGRID	377,595	40,739	4,912
	ChEMBL	2,642,425	138,464	127,353
	DrugBank	18,579	16	8
	GNP	3,354	-	-
	IntAct	297,422	65,154	6,973
	PID	11,285	-	-
	Reactome	21,080	14,463	11,049
Pathway	PID	318	-	-
	Reactome	2,171	1,408	1,284
Phenotype	OMIM	8,565	-	-

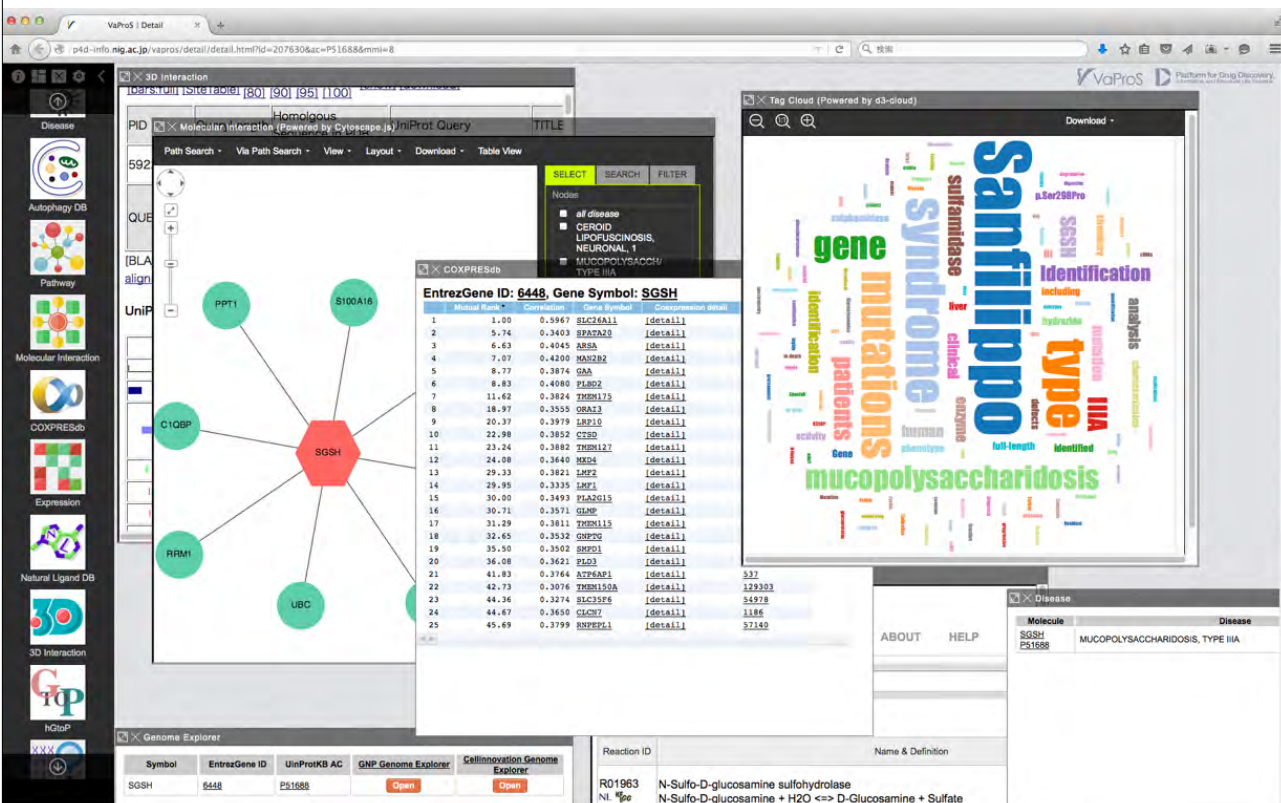
Data Type	DB	Count
Ligand	ChEMBL	1,623,526
	DrugBank	10,432

「計算生命科学の基礎V」

データクラウド：VaProS

[illegible]

ムコ多糖症を調べる

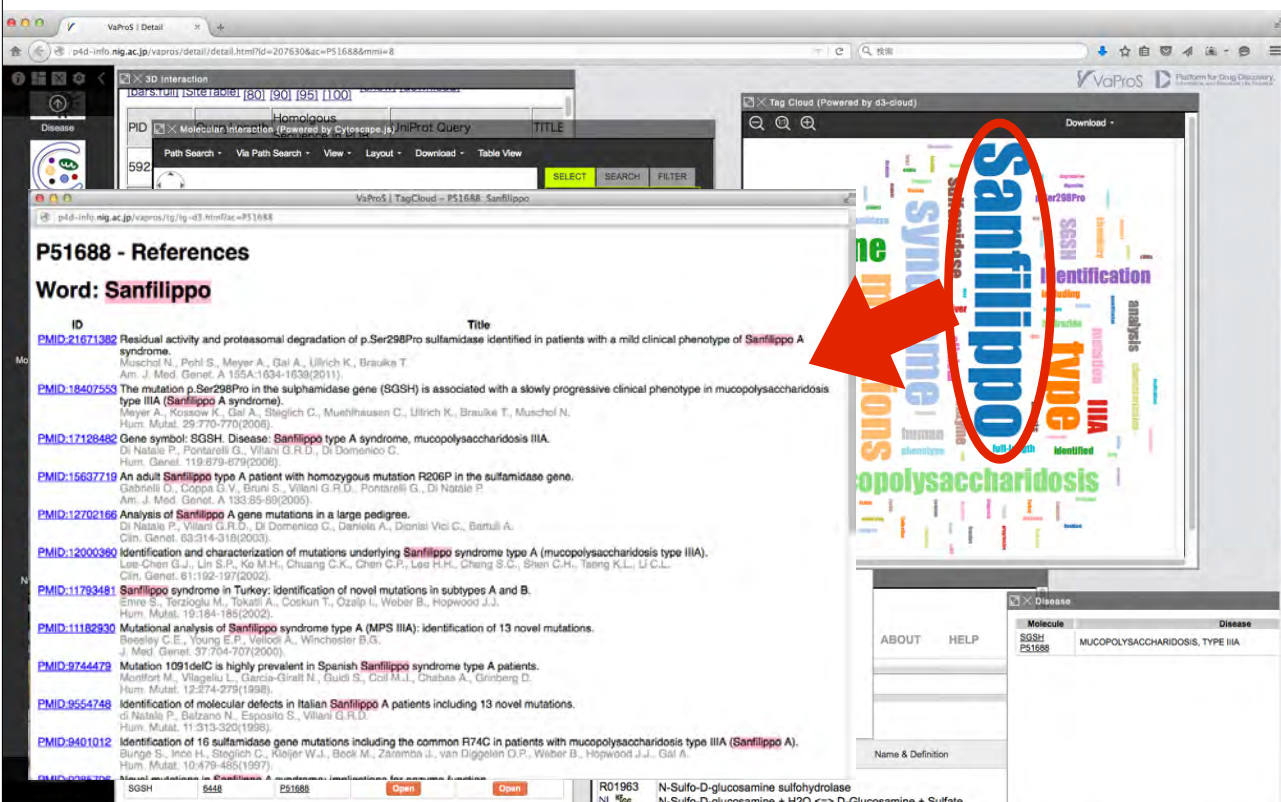


「計算生命科学の基礎V」

検索結果初期画面

29

ムコ多糖症を調べる

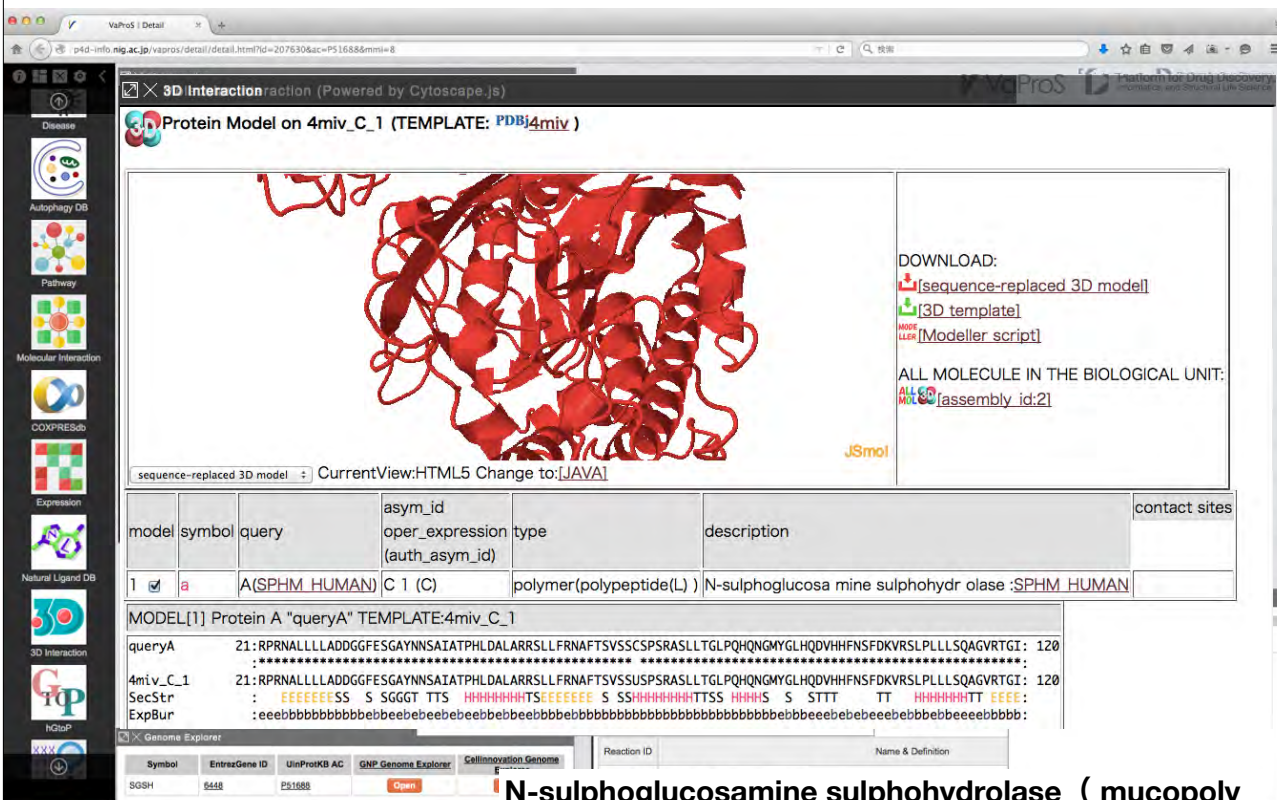


「計算生命科学の基礎V」

mucopolysaccharodosis type IIIA は、Sanfilippo type A syndromeともよばれている。

30

ムコ多糖症を調べる

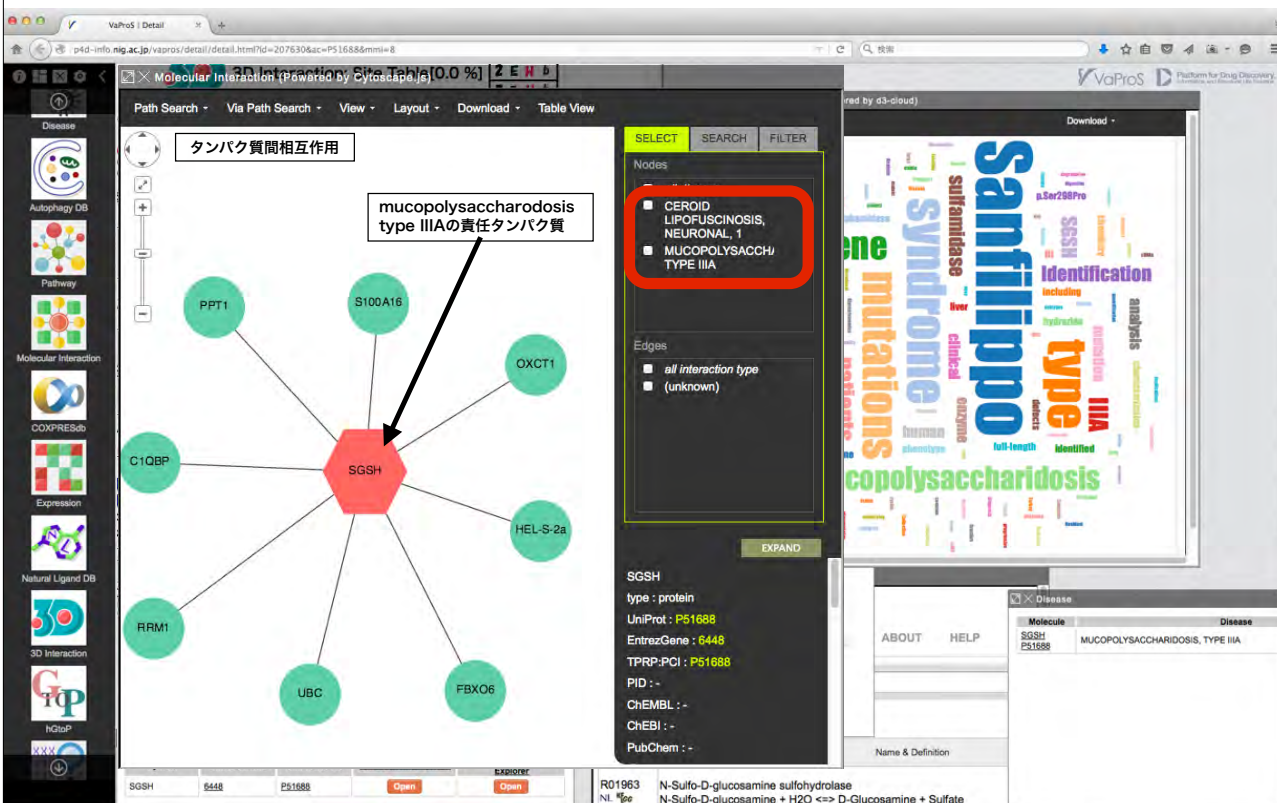


「計算生命科学の基礎V」

N-sulphoglucosamine sulphohydrolase (mucopoly saccharodosis type IIIAの責任タンパク質) は、立体構造が明らかになっている

31

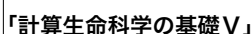
ムコ多糖症を調べる



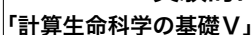
「計算生命科学の基礎Ⅴ」

実験的に相互作用が明らかになっているタンパク質群

32

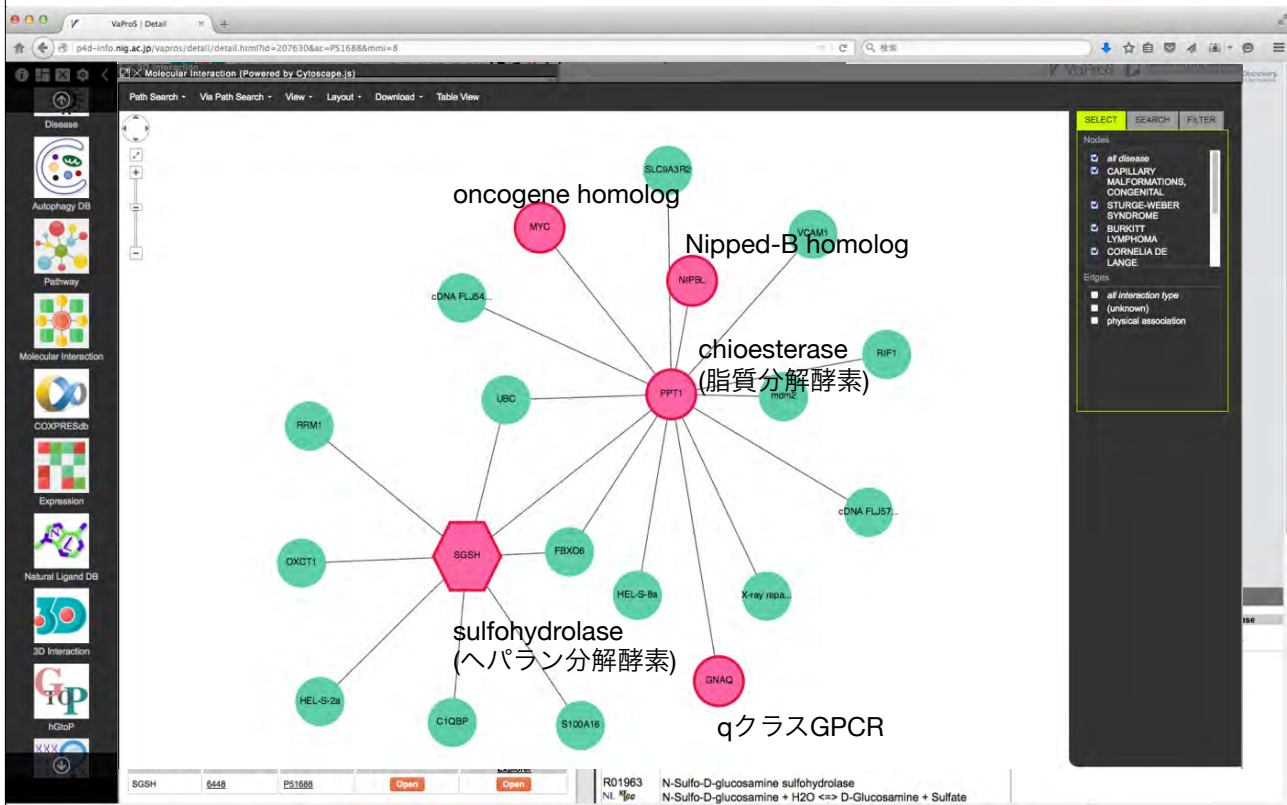


実験的に相互作用が明らかになっているタンパク質群



実験的に相互作用が明らかになっているタンパク質群

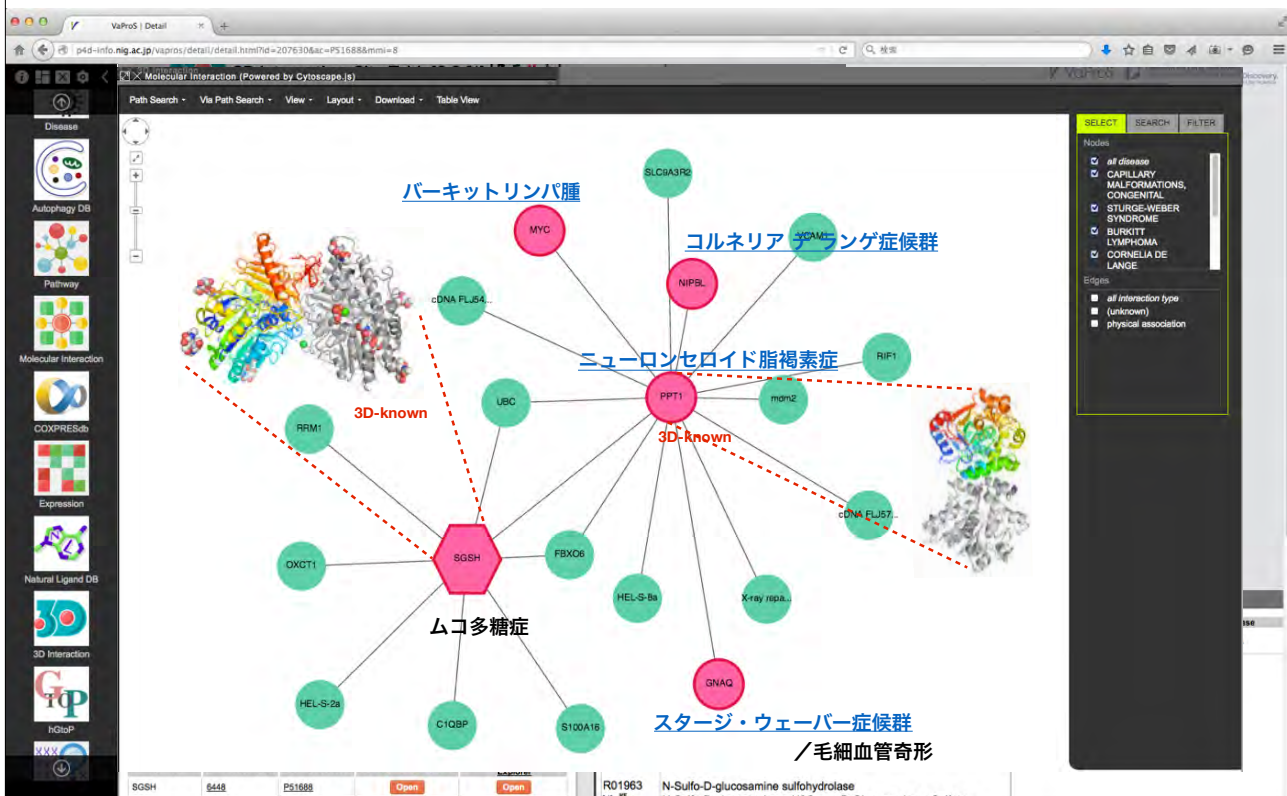
ムコ多糖症を調べる



実験的に相互作用が明らかになっているタンパク質群
「計算生命科学の基礎V」

35

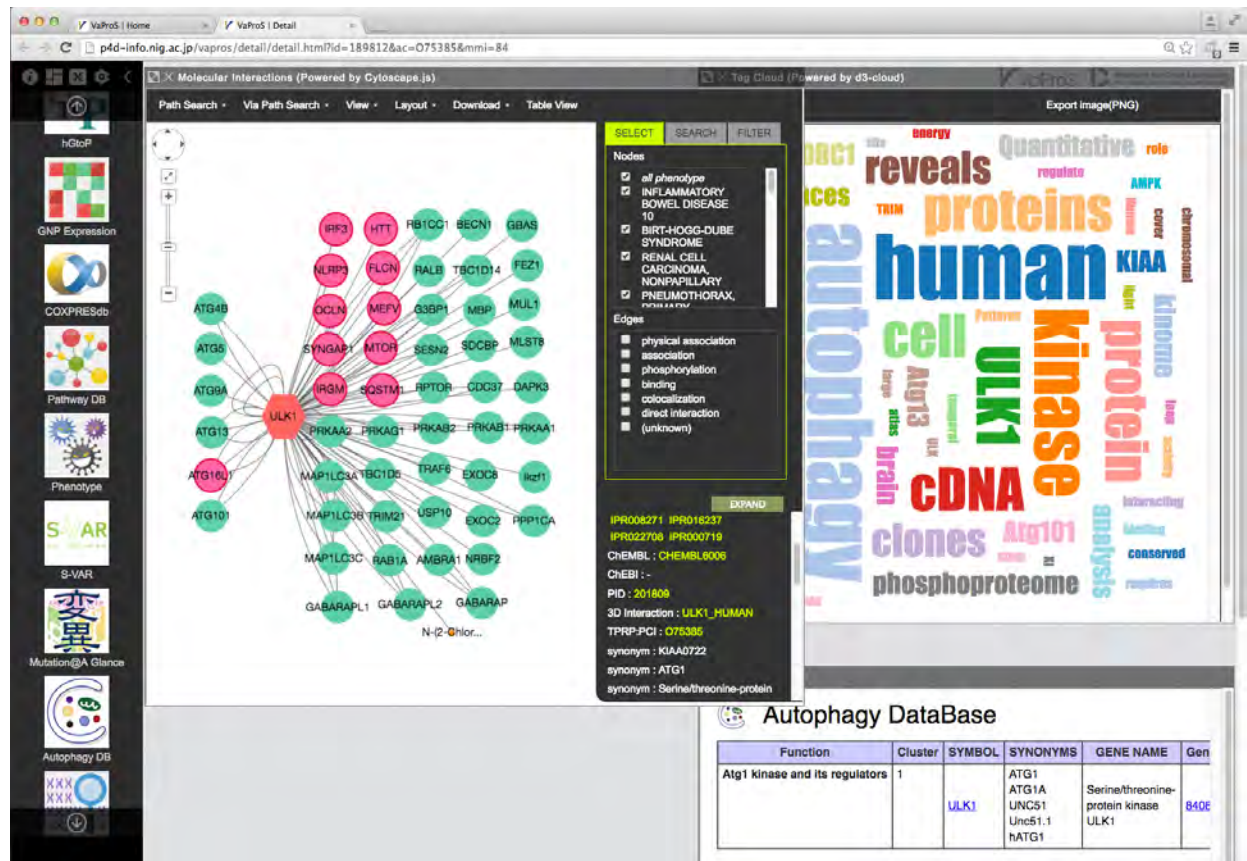
ムコ多糖症を調べる



異なる疾患がタンパク質相互作用を通じてつながっていることが、
生命科学ビッグデータからわかった。
「計算生命科学の基礎V」

36

オートファジー：ULK1を調べる

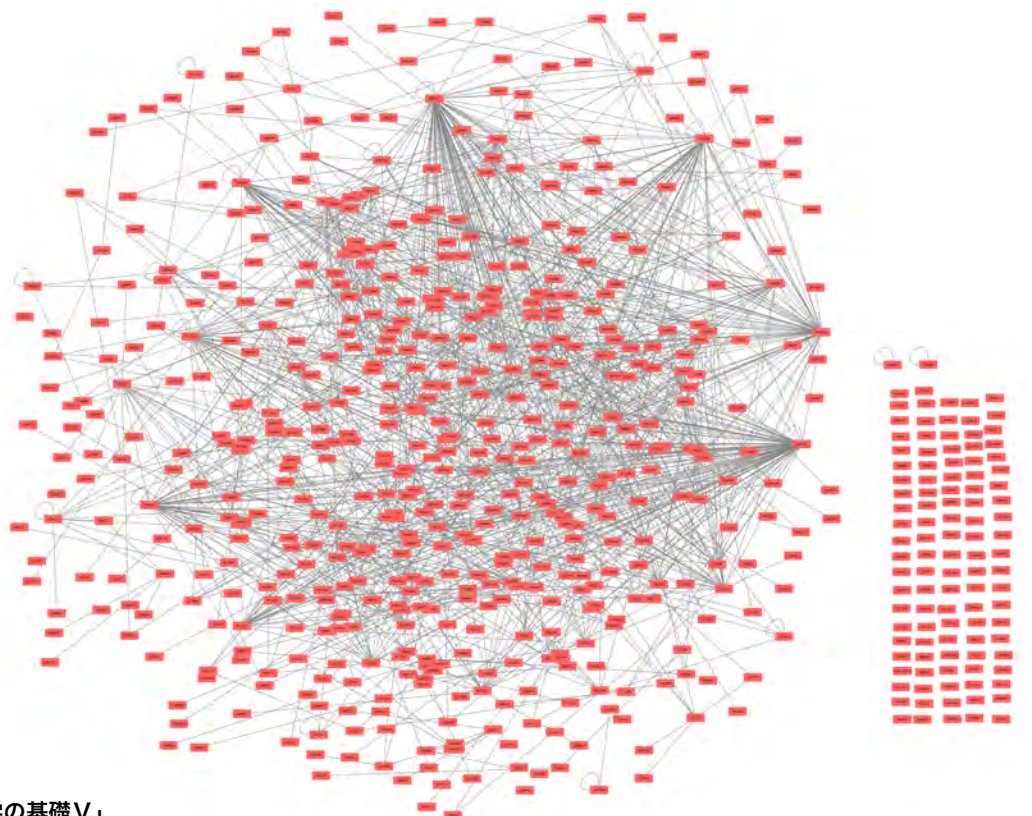


「計算生命科学の基礎V」

37

オートファジー

VaProSで、全相互作用データを取り出す



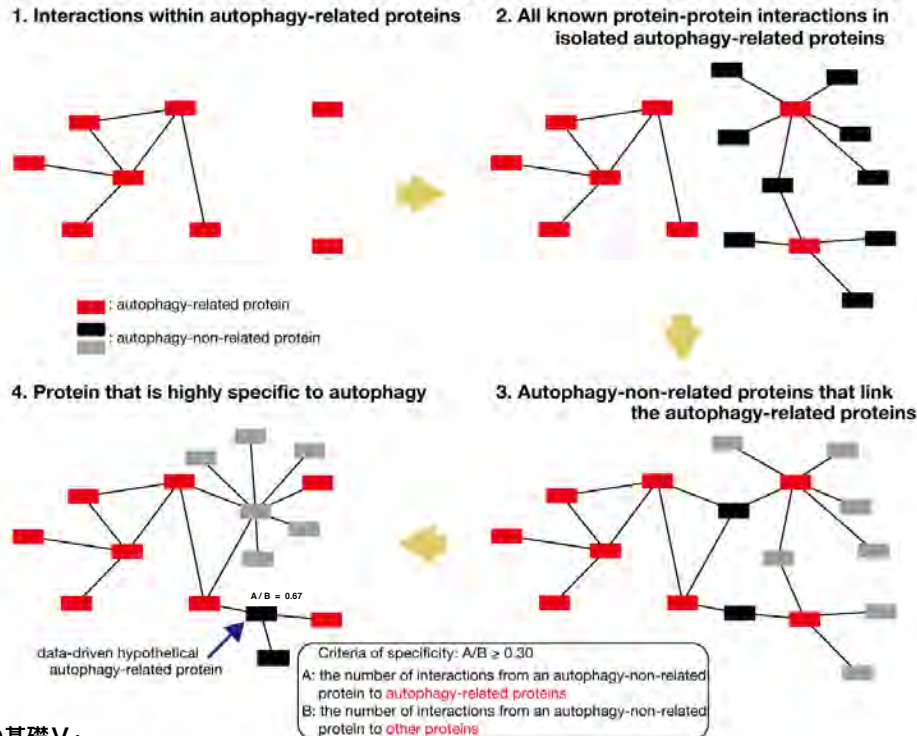
「計算生命科学の基礎V」

38

オートファジー

相互作用データから新規オートファジータンパク質を探す

Scheme of searching for candidate proteins on VaProS-Autophagy DB



「計算生命科学の基礎V」

39

オートファジー

相互作用データから新規オートファジータンパク質を探す

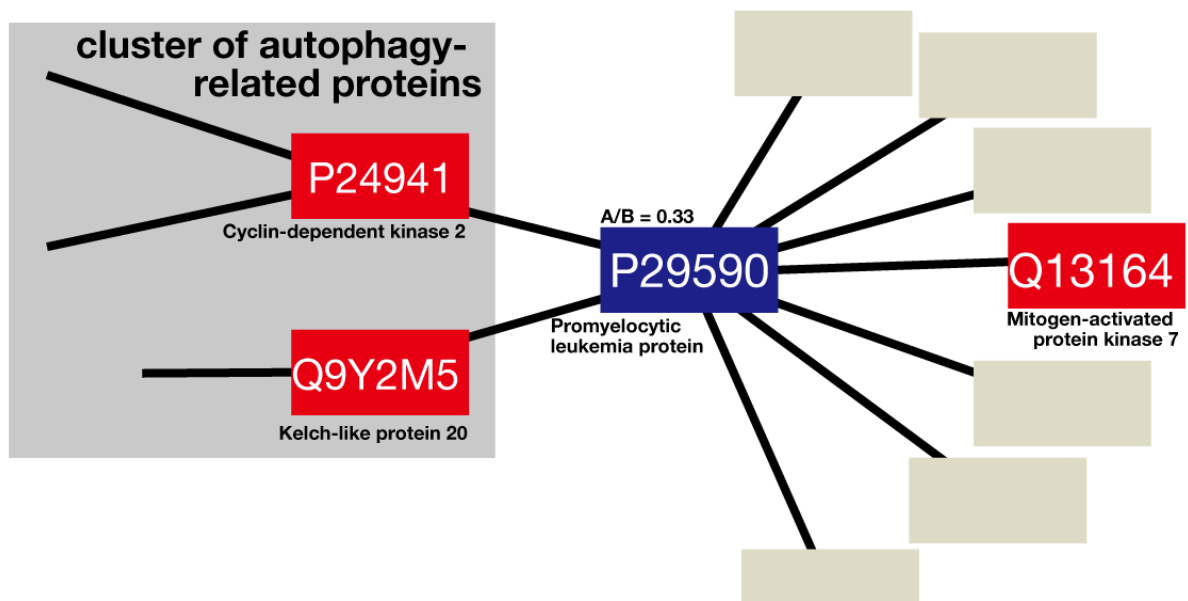


Table 1. The top five hits of the hypothetical autophagy-related proteins in humans.

#	Uniprot AC	A*	B*	A/B*	protein description	reference
1	B2RXCI	6	13	0.462	Trafficking protein particle complex subunit 11	Kim et al. 2016
2	P29320	2	5	0.400	Ephrin type-A receptor 3	Chukkapalli et al. 2014
3	P00519	15	40	0.375	Tyrosine-protein kinase ABL1	Yogalingam et al. 2008
4	P29590	3	9	0.333	Promyelocytic leukemia protein	Salmoni et al. 2012
5	P23727	4	13	0.308	Phosphatidylinositol 3-kinase regulatory subunit alpha	-

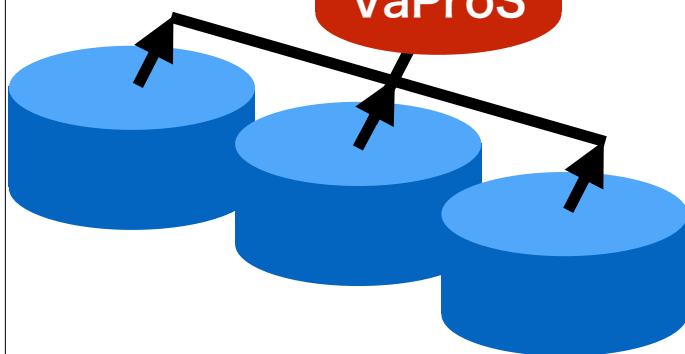
「計算生命科学の基礎V」

40

おわりに 知識ベースへの挑戦



生命科学の
問い合わせ



<https://en.wikipedia.org/wiki/LCARS>

データ間のルールが存在する

データが存在する

「計算生命科学の基礎V」

この仕組みを実現することをめざす