タンパク質構造予測とＨＰモデル

構造予測

実際のタンパク質 → 複雑、困難

↓ 簡略化

有用な
アルゴリズム ↑

ＨＰモデル → 取り組みやすい

構造予測

タンパク質を構成するアミノ酸

グリシン	Gly	バリン	Val
セリン	Ser	システイン	Cys
グルタミン	Gln	イソロイシン	Ile
フェニルアラニン	Phe	トリプトファン	Trp
グルタミン酸	Glu	リジン	Lys
アラニン	Ala	ロイシン	Leu
スレオニン	Thr	アスパラギン	Asn
チロシン	Tyr	メチオニン	Met
プロリン	Pro	アスパラギン酸	Asp
ヒスチジン	His	アルギニン	Arg

: 親水性　　: 疎水性

ＨＰモデルのルール

- アミノ酸をＨ（hydrophobic,疎水性、非極性アミノ酸）とＰ(polar,親水性、極性アミノ酸）のいずれかに分ける。
- Ｈは、水を嫌い、互いに引き付けあう
- ＨとＨが隣り合うと、ＨＨ結合が生まれる。ＨＨ結合はより低いエネルギーを取る。よってＨＨ結合が最も多い構造を最良とする。これをエネルギー最小化法という。

タンパク質→ＨＰモデル

例：

タンパク質：Lys-Val-Arg-Leu- Ile- Asp-Glu-Phe

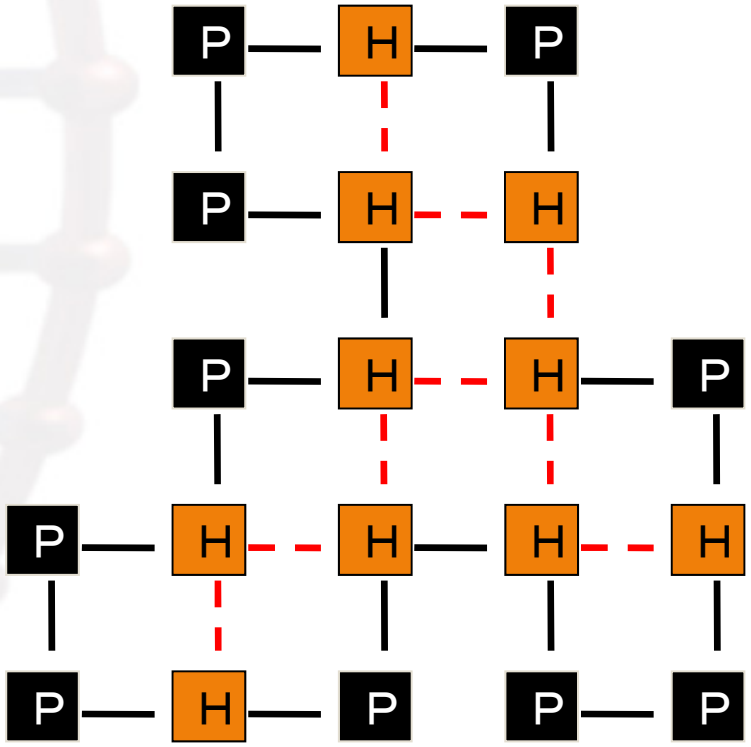
↓

ＨＰモデル：Ｈ -Ｐ- Ｈ- Ｐ - Ｐ－Ｈ - Ｈ -Ｐ

ＨＰモデル二次元格子構造の例

H-P-H-P-P-H-H-P-H-P-P-H-P-H-H-P-P-H-P-H

↓



左図では、ＨＨ結合は９個なので、

９点

点数が高い方が

良い結果である。

実験設定

- 配列は長さの違う11種類を用意する。
- 配列は全て最適解が既知のものである。
- ＧＡ，ＧＰ，ＡＣＯ，NNなどを用いる。
  - PERM（Pruned Enriched Rosenbluth Method）

実験結果

長さ	最適解
20	9
24	9
25	8
36	14
48	23
50	21
60	36
64	42

長さ	最適解
85	53
99	48
100	50

http://www.cs.ubc.ca/~hoos/Publ/ShmHoo03.pdf

に問題の記述がある

（ACOによる解法例の論文）

Seq. No.	*Length*	E^*	*Protein Sequence*
1	20	**-9**	$(HP)_2PH_2PHP_2HPH_2P_2HPH$
2	24	**-9**	$H_2P_2(HP_2)_6H_2$
3	25	**-8**	$P_2HP_2H_2P_4H_2P_4H_2P_4H_2$
4	36	-14	$P_3H_2P_2H_2P_5H_7P_2H_2P_4H_2P_2HP_2$
5	48	-23	$P_2HP_2H_2P_2H_2P_5H_{10}P_6H_2P_2H_2P_2HP_2H_5$
6	50	-21	$H_2(PH)_3PH_4PHP_3HP_3HP_4HP_3HP_3HPH_4(PH)_3PH_2$
7	60	-36	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$
8	64	-42	$H_{12}(PH)_2(P_2H_2)_2P_2H(P_2H_2)_2P_2H(P_2H_2)_2P_2HPHPH_{12}$
9	85	-53	$H_4P_4H_{12}P_6(H_{12}P_3)_3HP_2(H_2P_2)_2HPH$
10	100	-50	$P_3H_2P_2H_4P_2H_3(PH_2)_3H_2P_8H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2$ $PHP_2HPH_3P_6H_3$
11	100	-48	$P_6HPH_2P_5H_3PH_5PH_2(P_2H_2)_2PH_5PH_{10}PH_2PH_7P_{11}H_7P_2H$ $PH_3P_6HPHP_2$

Table 1. Benchmark instances for the 2D HP Protein Folding Problem used in this study with optimal or best known energy values E^*. (E^* values printed in bold-face are provably optimal.) The first eight instances can also be found at http://www.cs.sandia.gov/tech_reports/compbio/tortilla-hp-benchmarks.html, Sequence 9 is taken from [10], and the last two instances are taken from [14]. (H_i, P_i, and $(\ldots)_i$ indicate i-fold repetitions of the respective symbol or subsequence.)

実験結果

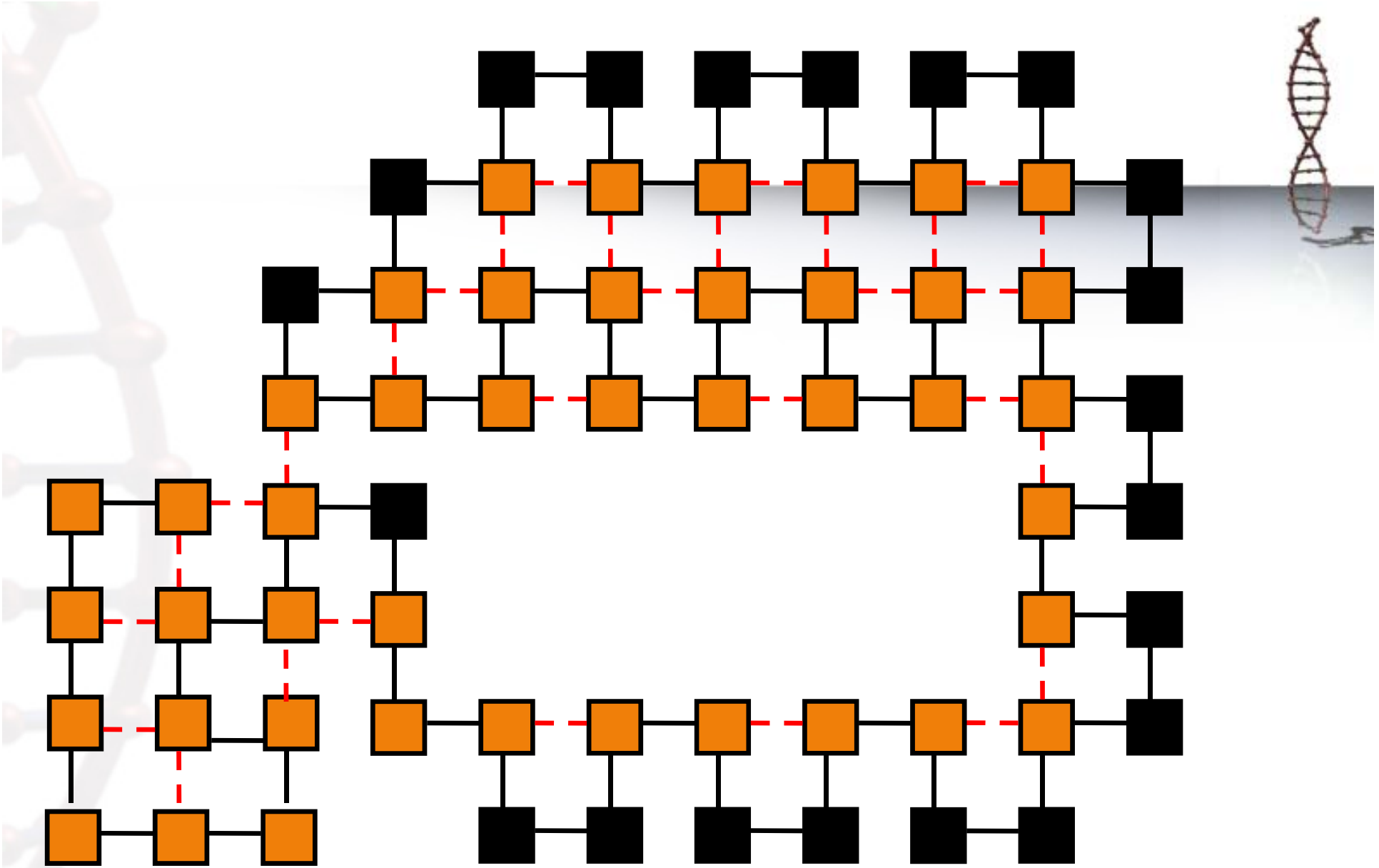
長さ	最適解	GP	GA	ACO	EMC	PERM
85	53	52		51	52	53
99	48			47		48
100	50			47		50

ACO（Ant Colony Optimization）

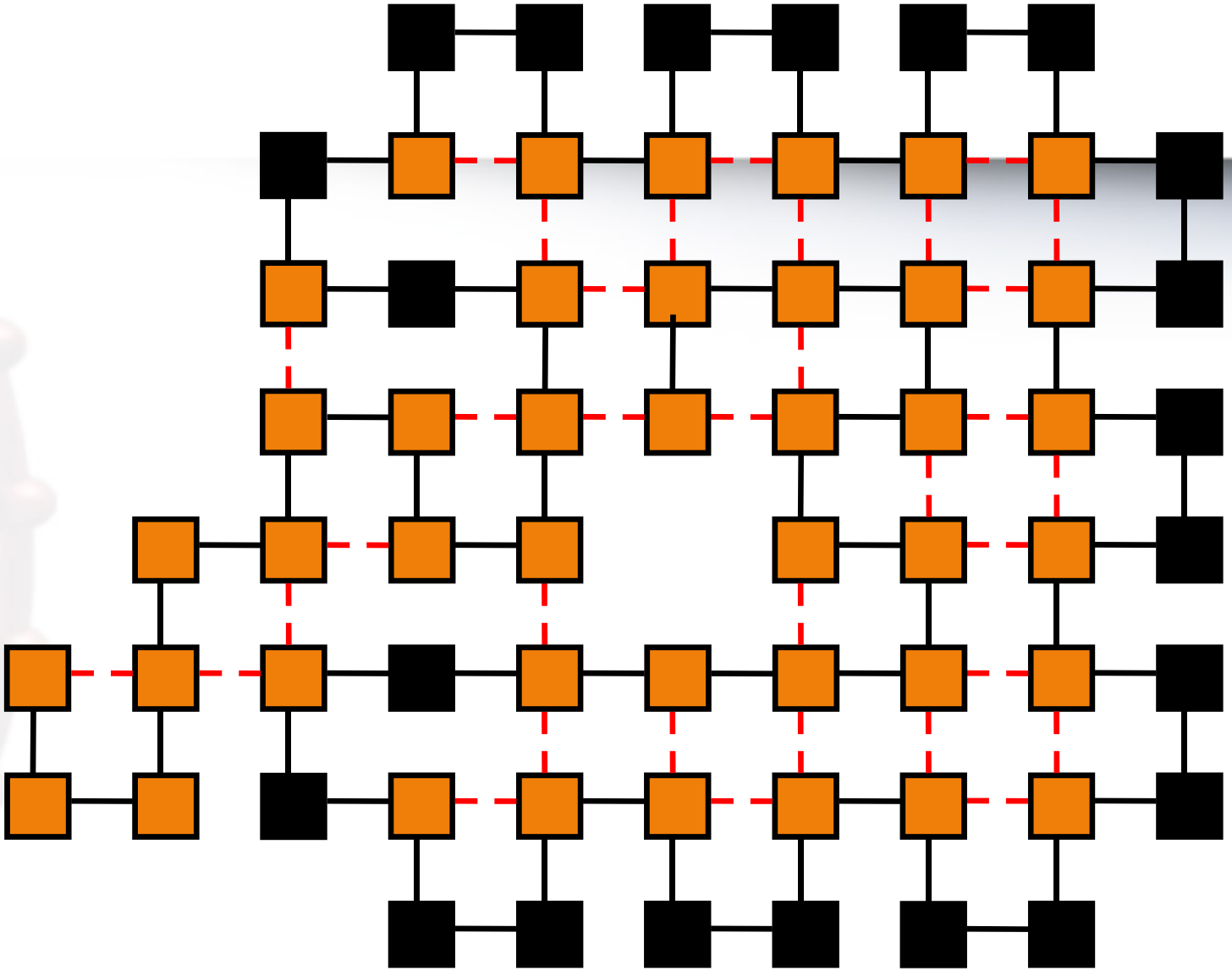
GA（Genetic Algorithm）

EMC（Evolutionary Monte Carlo）

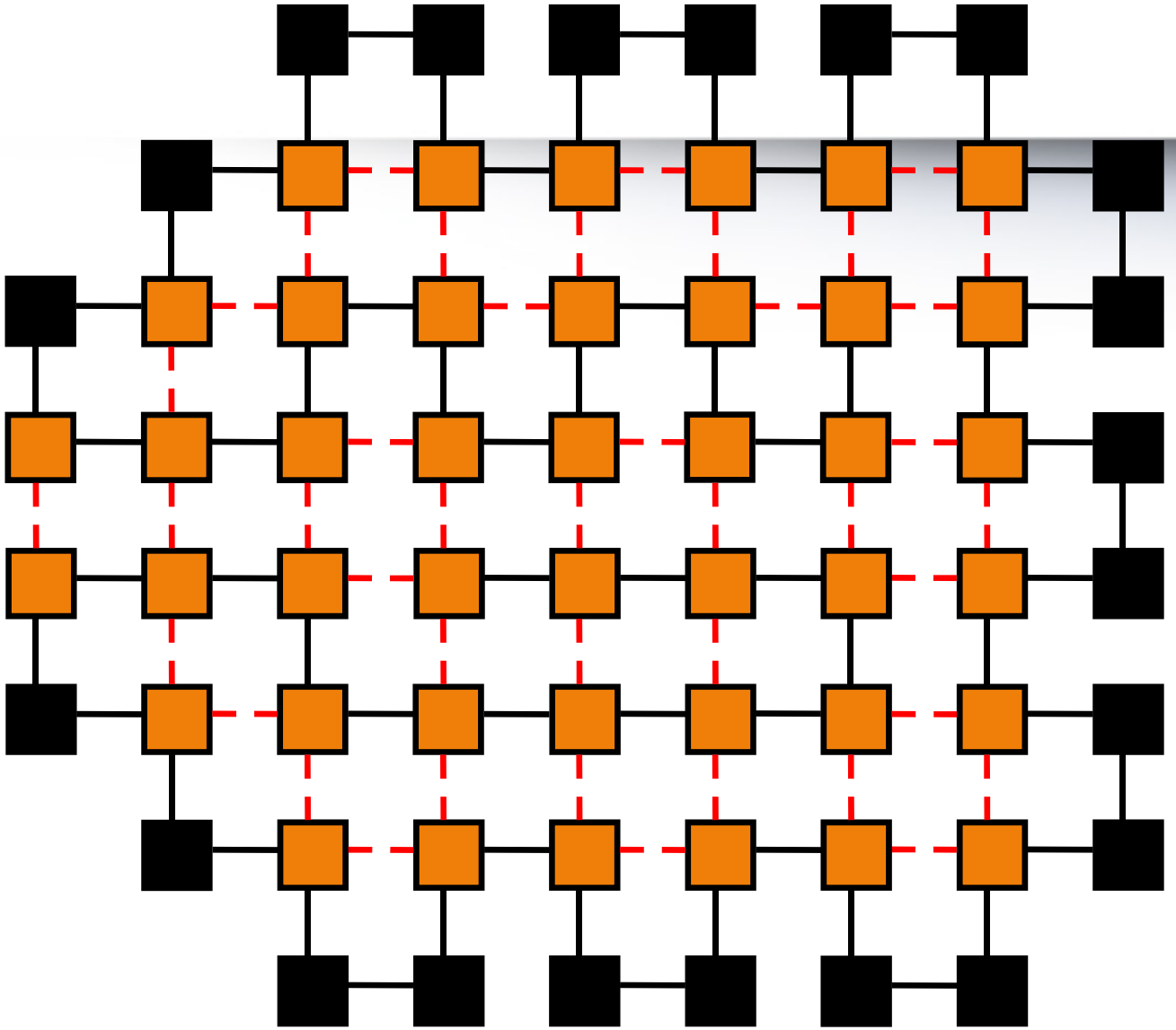
PERM（Pruned Enriched Rosenbluth Method）



１世代目（３０点）



８世代目（３４点）



１６代目（４２点、最適解）