Ninf-1/Ninf-Gを用いた NMR蛋白質立体構造決定のための 遺伝アルゴリズムのグリッド化

小野功(東工大),
水口尚亮, 中島直敏, 小野典彦(徳島大),
中田秀基(産総研 / 東工大), 松岡聡(東工大),
関口智嗣(産総研), 楯真一(生物分子工学研究所)

はじめに

蛋白質立体構造決定

- 核磁気共鳴法(NMR)は有望な解析手段の一つ
- データ解析のNOE帰属決定は人的・時間的コストが非常に高い！

・・・専門家が試行錯誤：数ヶ月～１年程度／個

GAによるNOE帰属の自動化の試み [Ono 02]

- 観測されたNOEを満たす立体構造探索問題として定式化
- 立体構造を遺伝アルゴリズム（Genetic Algorithm; GA）により最適化
- １３ 残基のα-helixで，専門家と同等の立体構造を自動的に求めることに成功
- ７８ 残基では，Pentium III 1.4GHz/1CPUで**約200日**かかる

⇒ ***高速化が緊急の課題***

- 現実的な時間での計算完了には，**数百～1,000CPU程度**での並列計算が必要

⇒ ***１研究室の計算資源では困難***

グリッド

- 複数拠点の計算資源を相互接続して大規模計算を可能にする次世代並列計算プラットフォームとして注目されている

目的

NMR蛋白質立体構造決定のための1,000CPU程度のグリッド向けGAシステムの提案とその有効性の検証

NMR蛋白質立体構造決定のための遺伝アルゴリズム [Ono 02]（1）

基本的な考え方

×NOE帰属 ⟶ 立体構造計算 （従来法）

○立体構造生成 ⟶ NOE帰属による評価 （本研究）

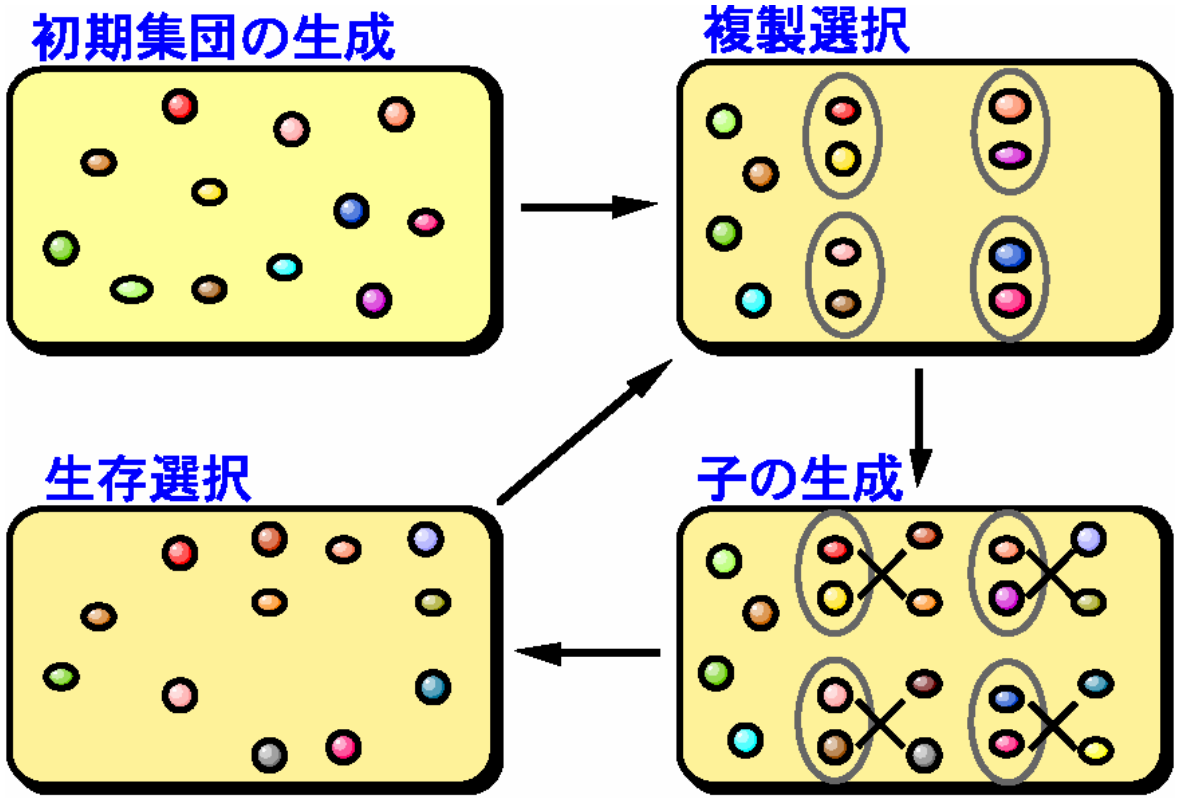
観測されたNOEシグナルをなるべくよく説明する立体構造を探索

アルゴリズムの概略

1. GAにおいて，解候補となる立体構造を生成.
2. 立体構造からNOEシグナルを予測.
3. 予測シグナルを観測されたNOEシグナルに帰属.
4. 帰属に成功した観測NOEの数に基づき評価値を計算.

NMR蛋白質立体構造決定のための遺伝アルゴリズム [Ono 02]（2）

遺伝的アルゴリズム（GA）の枠組みと設計項目



- コード化/交叉・突然変異設計 …… 解表現，子の生成方法
- 世代交代モデル設計 …… 複製/生存選択方法

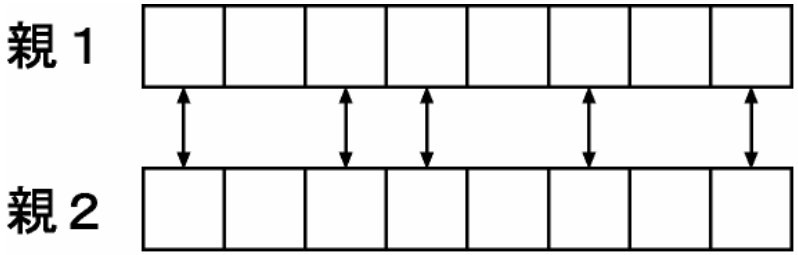
NMR蛋白質立体構造決定のための遺伝アルゴリズム [Ono 02]（3）

コード化

- 二面角(ϕ, ψ, ω, χ)からなる実数ベクトル

交叉

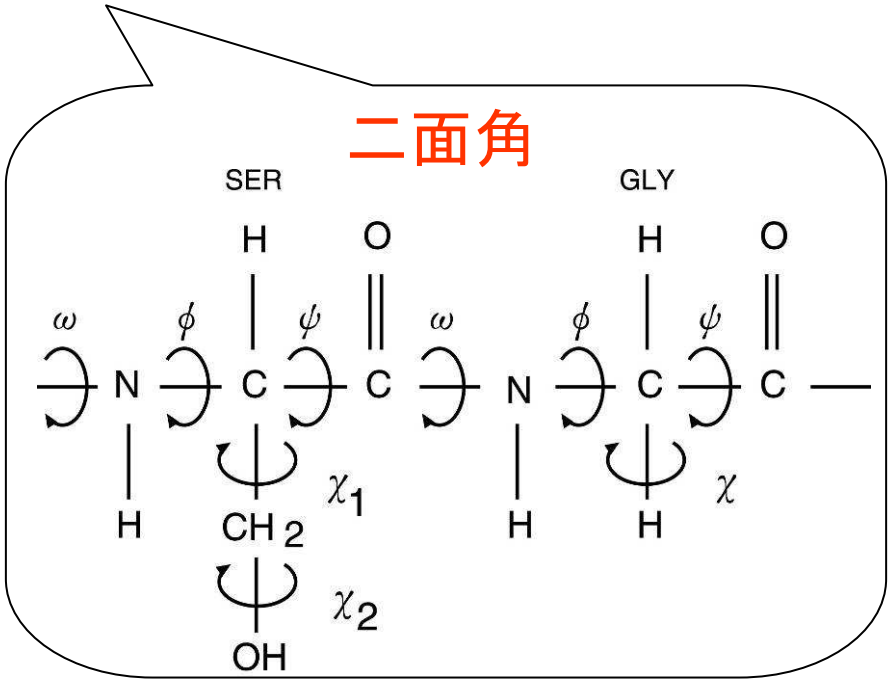
- 一様交叉



各要素５０%の確率で入れ替え

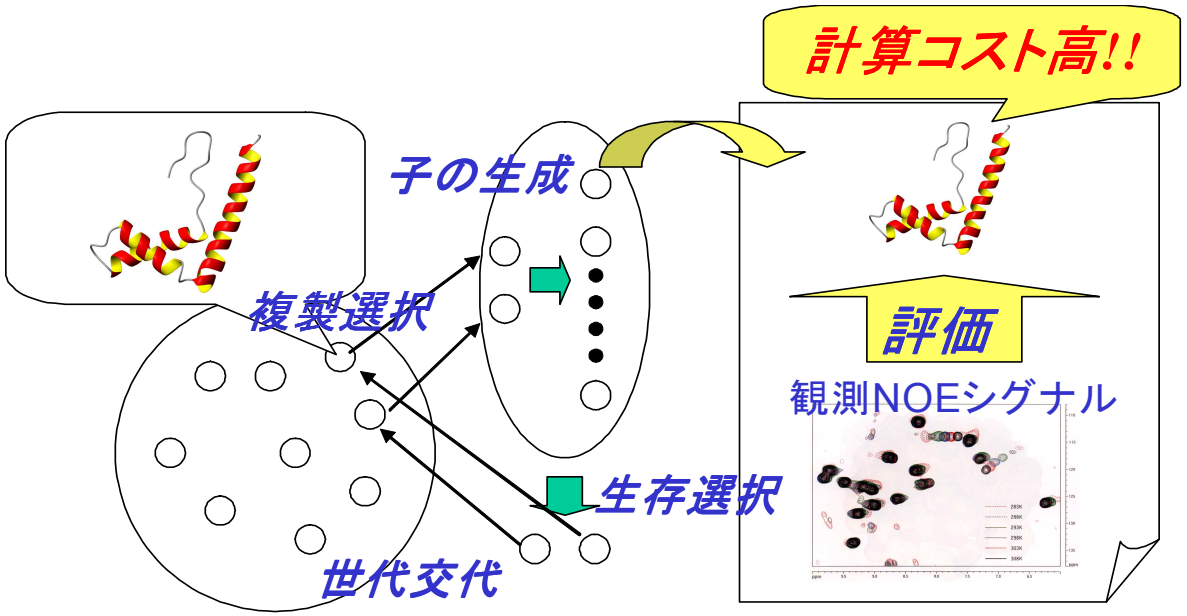
突然変異

- 各要素1%の確率で[-1°, +1°]の一様乱数を加える

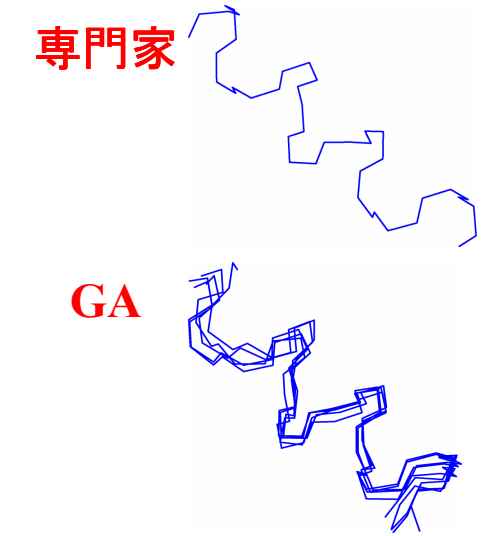


NMR蛋白質立体構造決定のための遺伝アルゴリズム [Ono 02]（4）

世代交代モデル：MGG [佐藤 97]



13残基のα-helix構造決定問題において，専門家に迫る構造を自動的に求めることに成功!!



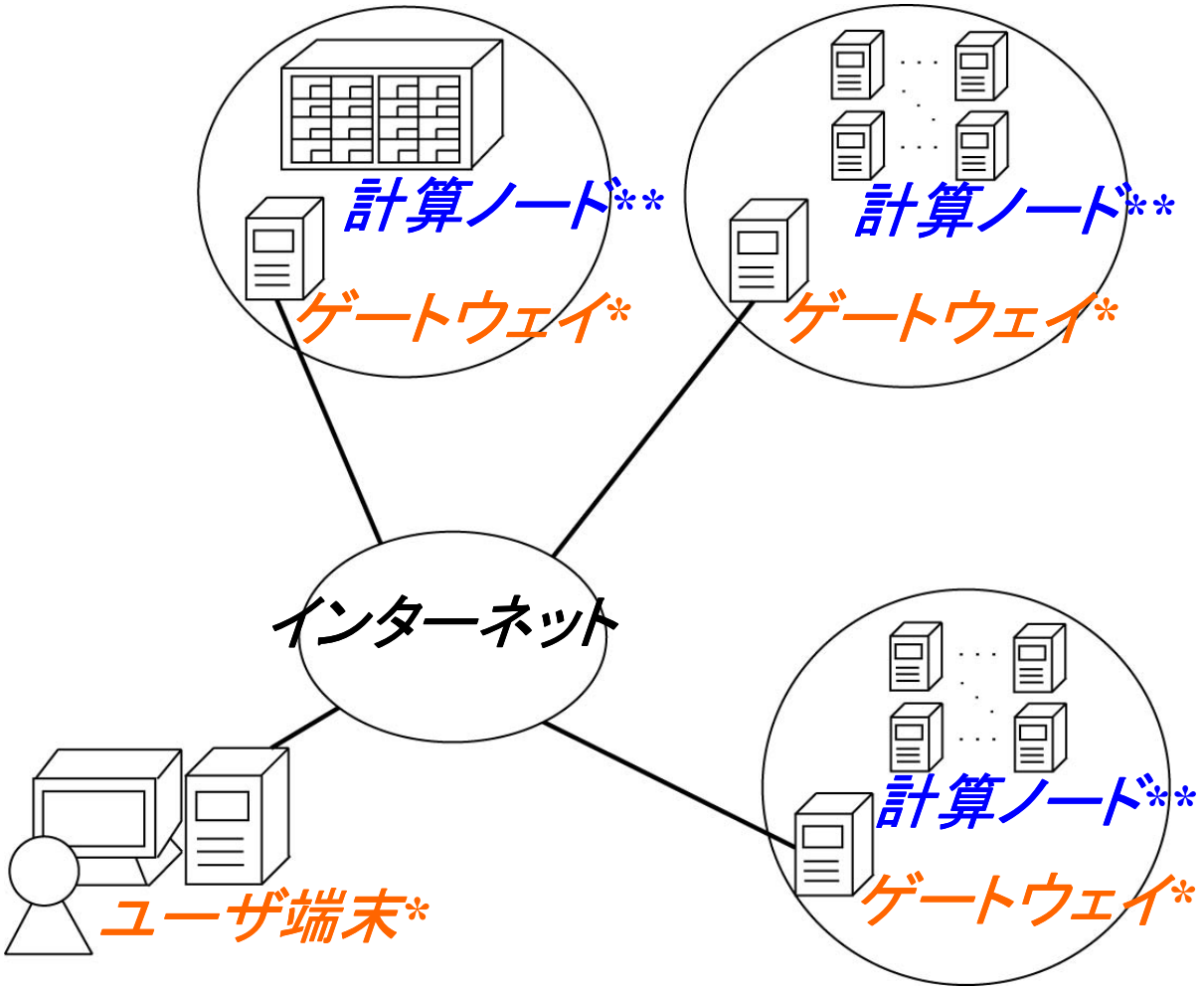
大規模立体構造決定での問題点

78 残基立体構造決定では，Pentium III 1.4GHz/1CPUで**約200日**かかる

現実的な時間での計算完了には，**数百～1000CPU**程度での並列計算が必要

高速化のためグリッド向けに並列化

想定するグリッド環境



*グローバルIP
**プライベートIP

NMR蛋白質立体構造決定のための グリッド向けGAシステムの解決すべき要件

スケーラビリティ:
1,000CPU程度までのスケーラビリティ

セキュリティ:
強力な認証機構，通信路の暗号化

耐障害性:
一部の計算ノードで障害が起こっても，全体の計算は継続される

ヘテロな環境への対応:
計算ノードの能力に応じてタスクを配分する仕掛けが必要

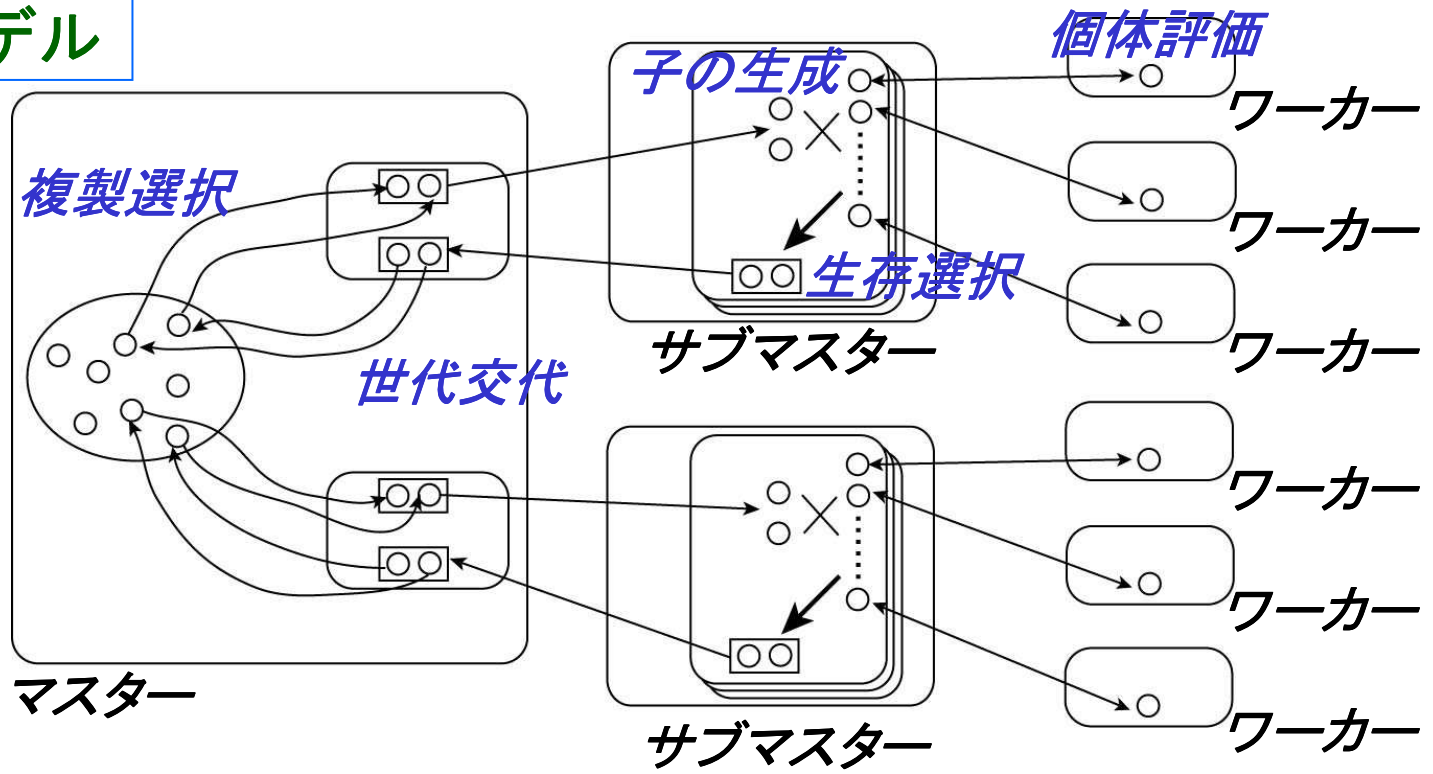
NATへの対応:
PCクラスタ内部のプライベートIPをもつ計算ノードをユーザ端末から利用できる必要がある

上記要件を解決したGAシステムの提案

NMR蛋白質立体構造決定のためのグリッド向けGAシステムの提案(1)

グリッド上での並列GAモデルの提案

提案モデル



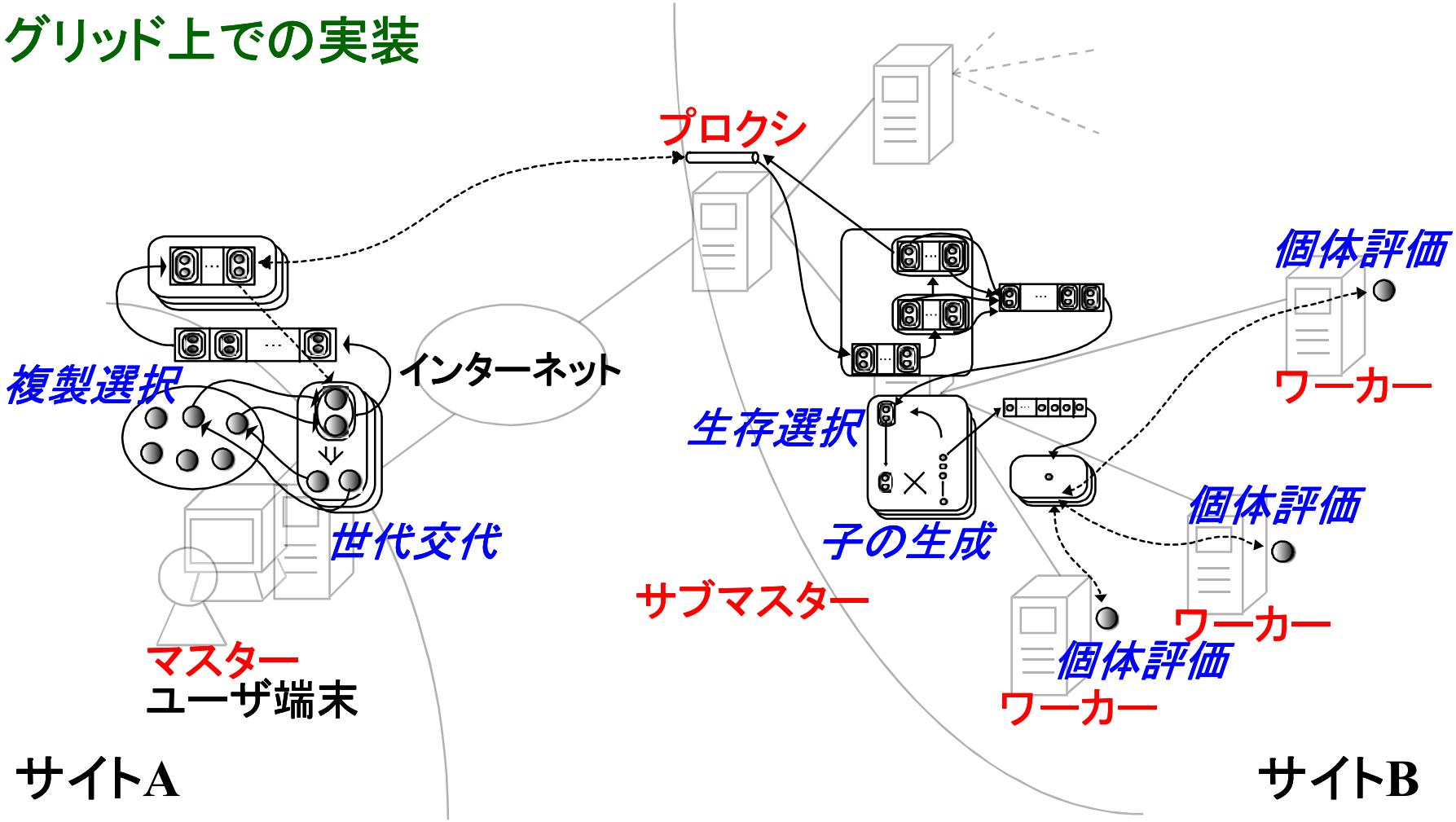
個体評価の計算時間コストが高く，独立に計算可能 ⇒ *評価計算を並列化*

1,000CPU程度のスケーラビリティに対応

- 世代交代ループを複数同時に実行 ⇒ *並列度向上*
- マスターの機能をマスターと複数のサブマスターに分散 ⇒ *負荷分散*

NMR蛋白質立体構造決定のためのグリッド向けGAシステムの提案(2)

グリッド上での実装



スケーラビリティに対応

・マスターをユーザ端末，サブマスターを他サイトに設置 ⇨ *通信量を抑える*

NMR蛋白質立体構造決定のためのグリッド向けGAシステムの提案(3)

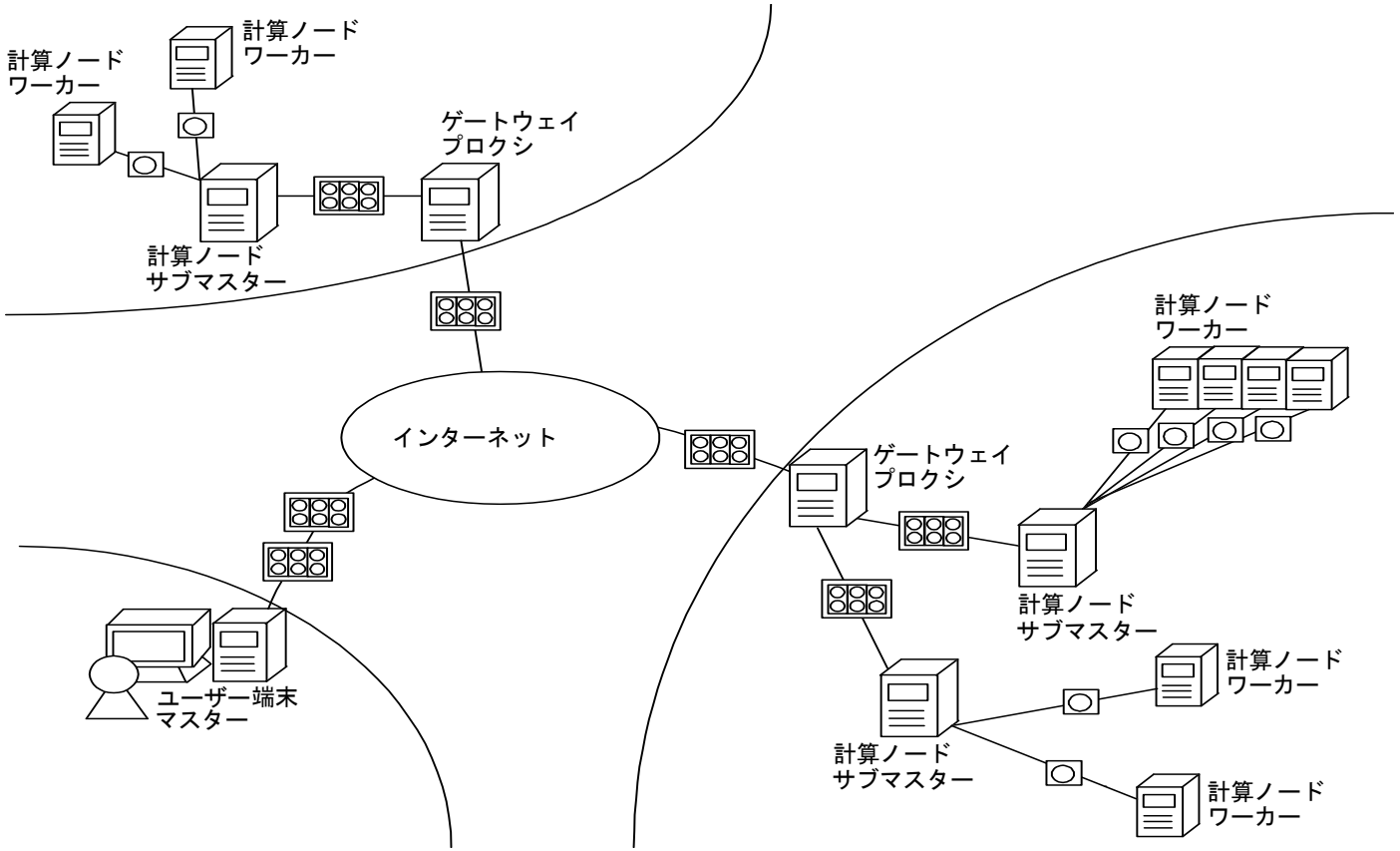
グリッド上での並列化の工夫(1)

スケーラビリティへの対応

・インターネットを介した通信はデータをまとめて回数を減す

・クラスタ内部の通信は，高速な **Ninf-1** を利用 ⇒ *通信遅延を抑える*

Ninf-1・・・並列処理支援のRPCミドルウェア，認証なし，通信の暗号化なし



NMR蛋白質立体構造決定のためのグリッド向けGAシステムの提案(4)

グリッド上での並列化の工夫(2)

セキュリティへの対応

・マスター・サブマスター間に **Ninf-G** を利用 ⇨ ***認証・暗号化***

Ninf-G・・・Ninf-1をグリッド上に拡張実装したGridRPCミドルウェア
グリッドのデファクトであるGlobus Took Kit [http://www.globus.org]
を利用し，認証，通信データの暗号化を提供

障害への対応

・障害箇所をスレッド・レベルで切り離し，全体が停止するのを防ぐ

ヘテロな環境への対応

・ワーカーのCPU性能により，サブマスター単位でクラスタリング

NATへの対応

・ゲートウェイ上で，通信を仲介するプロクシを動作

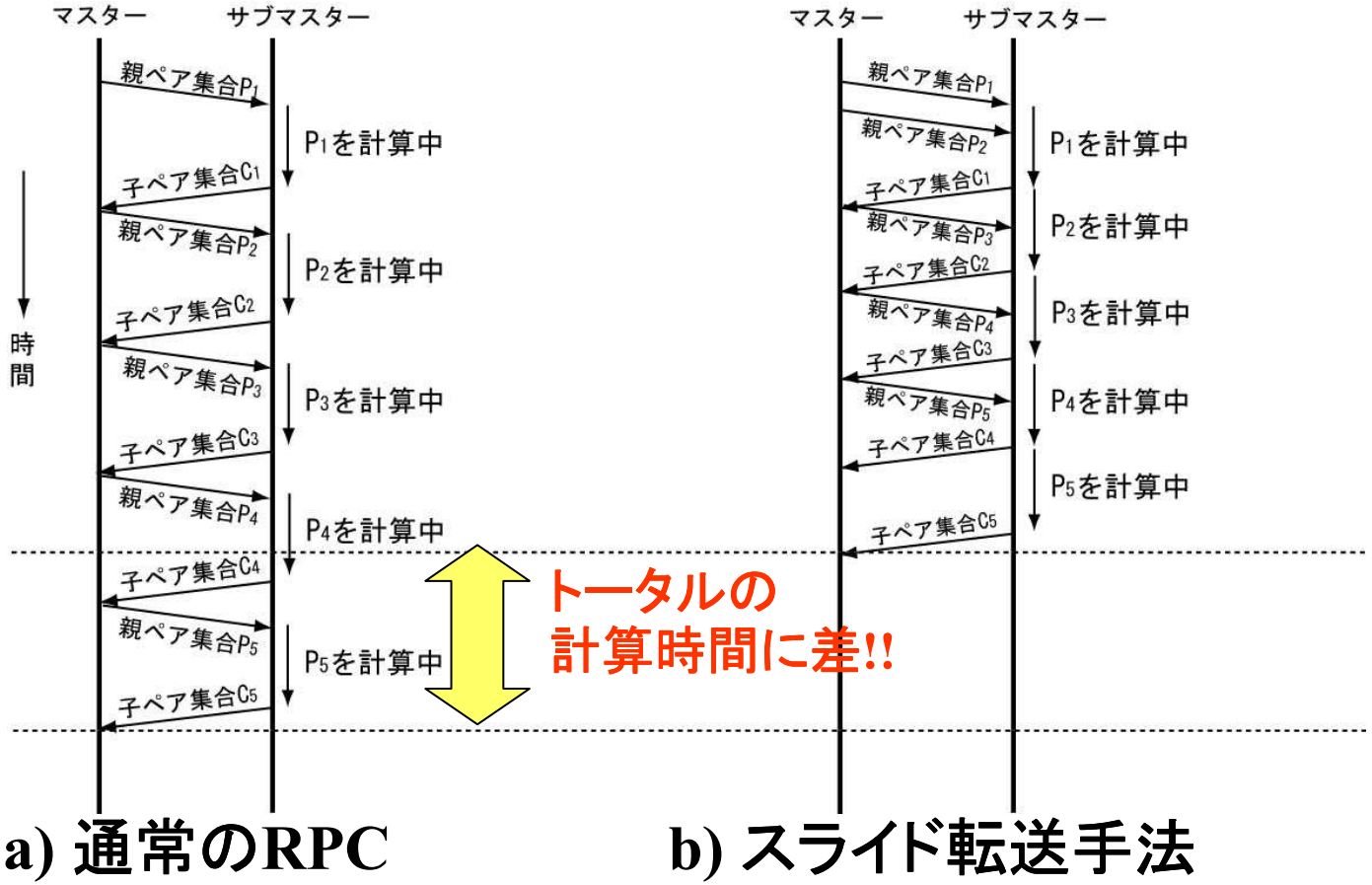
NMR蛋白質立体構造決定のためのグリッド向けGAシステムの提案(5)

スライド転送方式による通信遅延の隠蔽

・マスターとサブマスター間は通信遅延が非常に大きい

⇐ 暗号／復号化，低速なインターネット上の通信

⇒ 通信遅延を見かけ上なくす工夫の導入



実験(1)

実験目的

- 提案システムのスケーラビリティ，スライド転送手法の有効性，耐障害性の検証
- 実運用を想定した運用試験

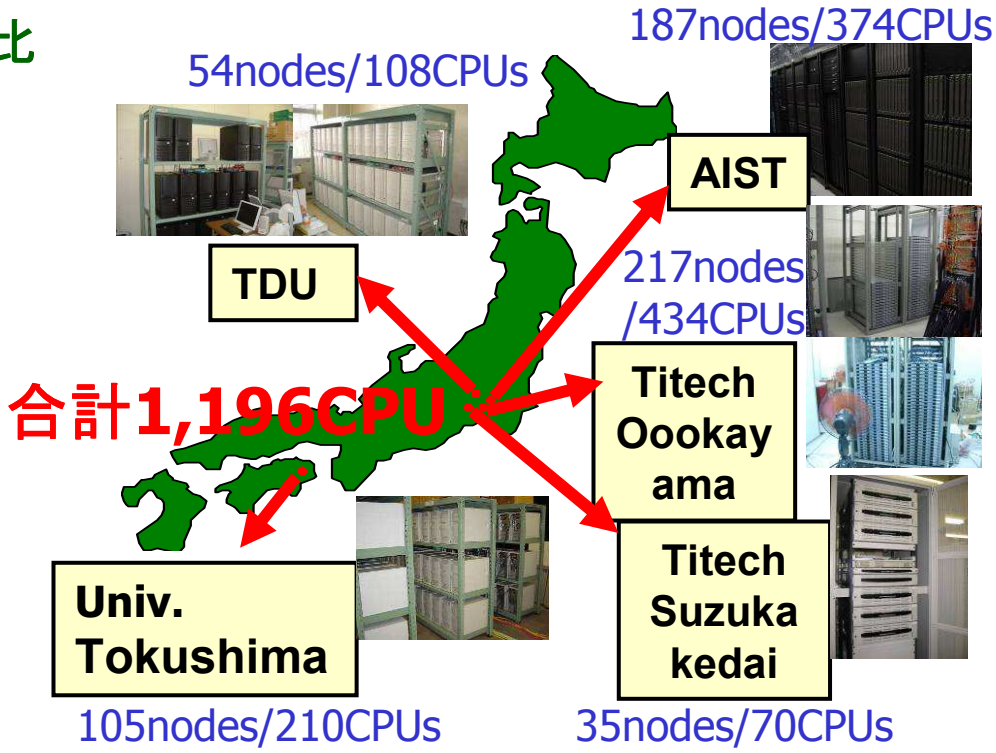
共通設定

- テストベッド：1196CPU/5サイト
- 対象問題：78残基hmg2b蛋白質
- Pentium III 1.4GHz での1個体の平均評価時間：2922.6529 [ms]
- 1個体のデータサイズ：約11.5KB

1個体の平均評価時間から求めた性能比

サイト名	CPU	CPU数	性能比
小野研	Athlon MP 2000+	126	1.1950
	Athlon MP 2800+	84	1.6588
合田研	Pentium III 1.4GHz	70	1.0000
藤沢研	Athlon MP 2400+	78	1.3415
	Opteron 240 (64bit)	30	1.3328
松岡研	Athlon MP 1900+	194	1.1734
	Athlon MP 2000+	66	1.2111
	Opteron 242	174	1.1661
産総研	Xeon 3.06GHz	374	2.2371

ACT-JSTグリッド



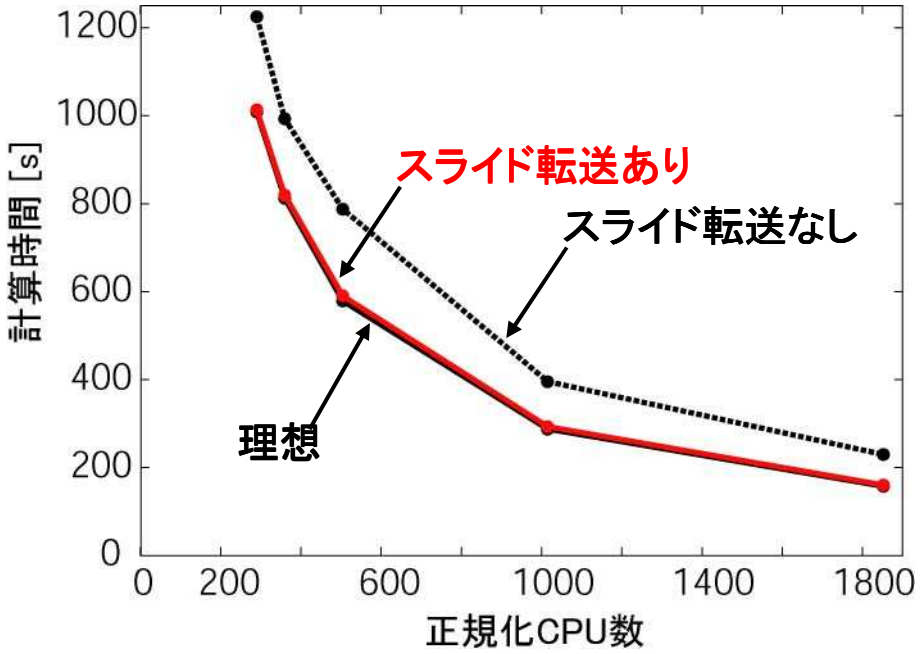
実験（2）

スケーラビリティの実験設定

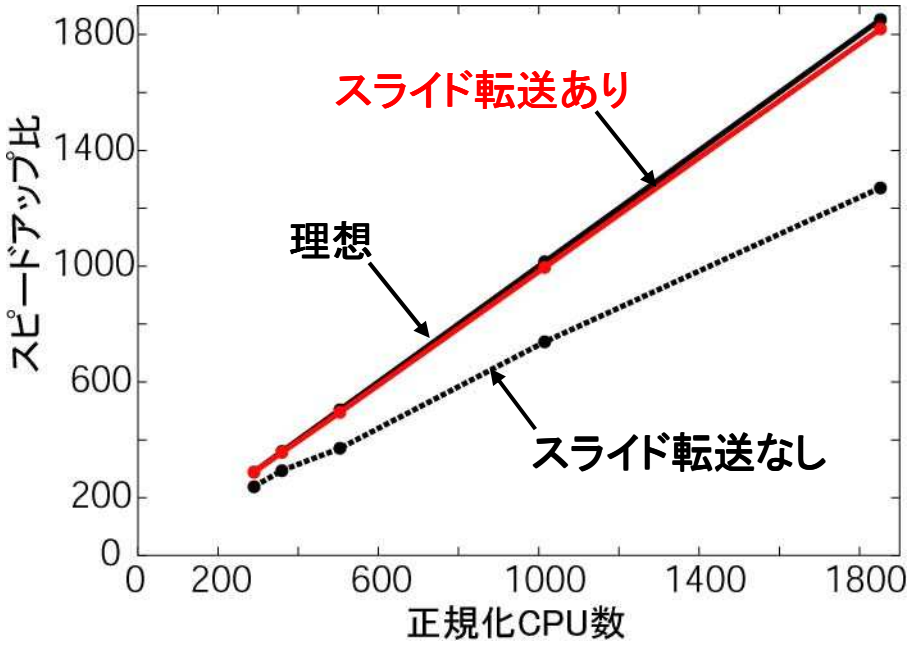
試行回数：独立3試行の平均　打ち切り世代数：500　生成子個体数／1世代：200
集団サイズ：500　世代交代スレッド数：10, 14, 16, 36, 50/サイト数で増加
正規化CPU数：Pentium III 1.4GHz／1CPUの平均評価性能を1.0として正規化

実験結果

CPU数 vs 計算時間



CPU数 vs スピードアップ比



1,196CPUで提案システムが
理想に近いリニアなスケーラビリティ!!

実験（3）

耐障害性の検証

以下のエラー状況を人為的に発生させた

- マスター⇔ゲートウェイ，ゲートウェイ⇔サブマスター，
  サブマスター⇔ワーカーの通信路が物理的に切断
- マスター⇔ゲートウェイ，ゲートウェイ⇔サブマスター，
  サブマスター⇔ワーカーのコネクションが切断
- ゲートウェイ，サブマスター，ワーカーのプログラムが異常終了

→ システム全体の処理は止まらずに
正常に計算を続けることを確認

実運用を想定した運用試験

CPU数：1,196（5サイト）　打ち切り世代数：30000　試行回数：独立3試行の平均

約2時間40分で正常に終了
（Pentium III 1.4GHz／1CPUで約200日かかる）

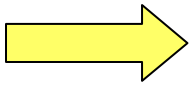
→ 実際の規模の問題を現実的な
時間内に計算を完了！！

おわりに（1）

まとめ

- Ninf-1/Ninf-Gを用いたNMR蛋白質立体構造決定のための
  グリッド向けGAシステムの提案
- 5サイト／1196CPUから構成されるグリッドテストベッド上で
  - 提案システムの理想的なスケーラビリティ
  - 耐障害性
  - 実際の規模の問題を現実的な時間内に計算完了可能

  を確認した



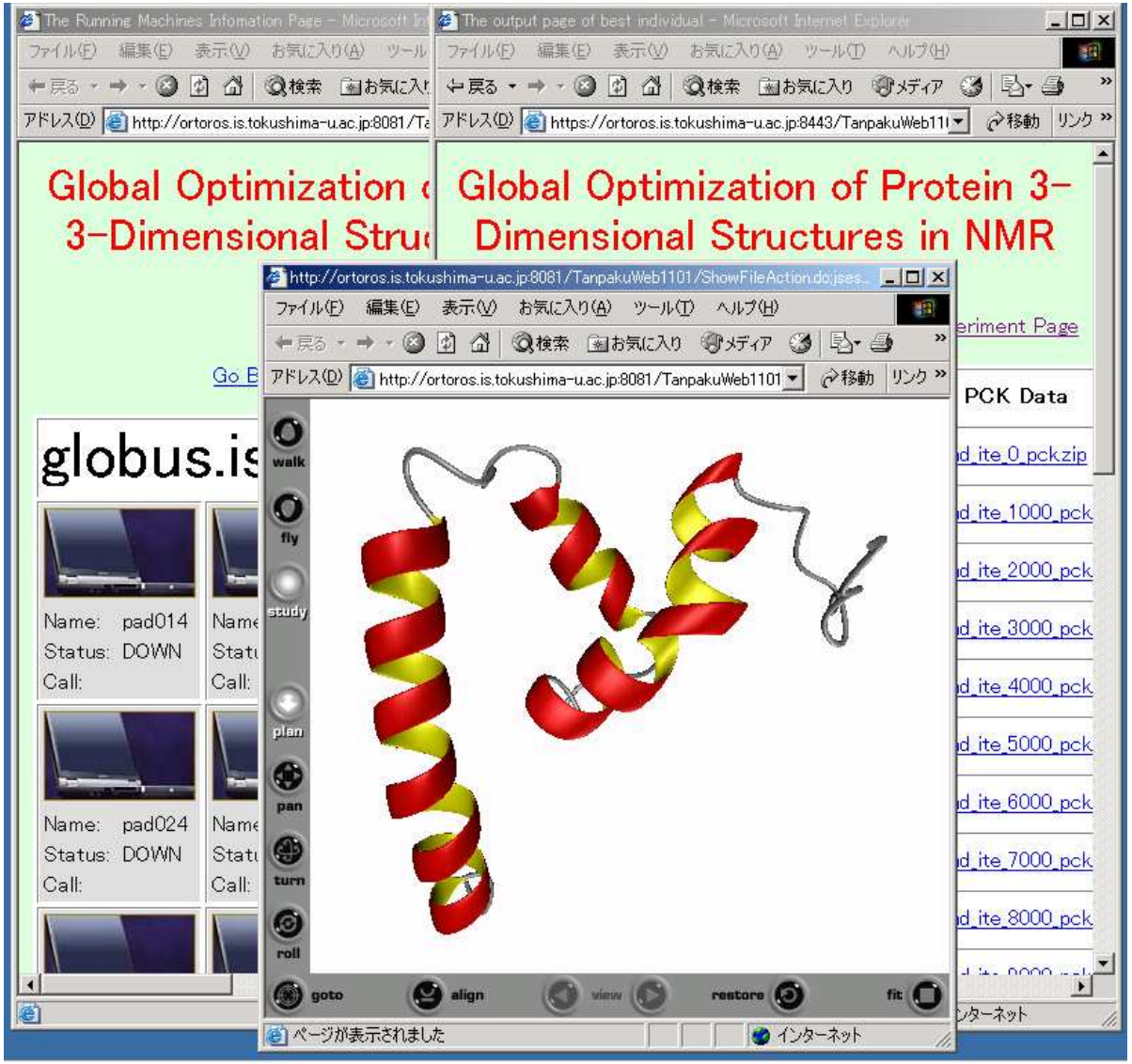
提案システムが有効であることを確認

今後の課題

- 専門家向けのグリッドポータルシステムの構築
- グリッド向けGAフレームワークの構築
- 探索アルゴリズムの性能向上

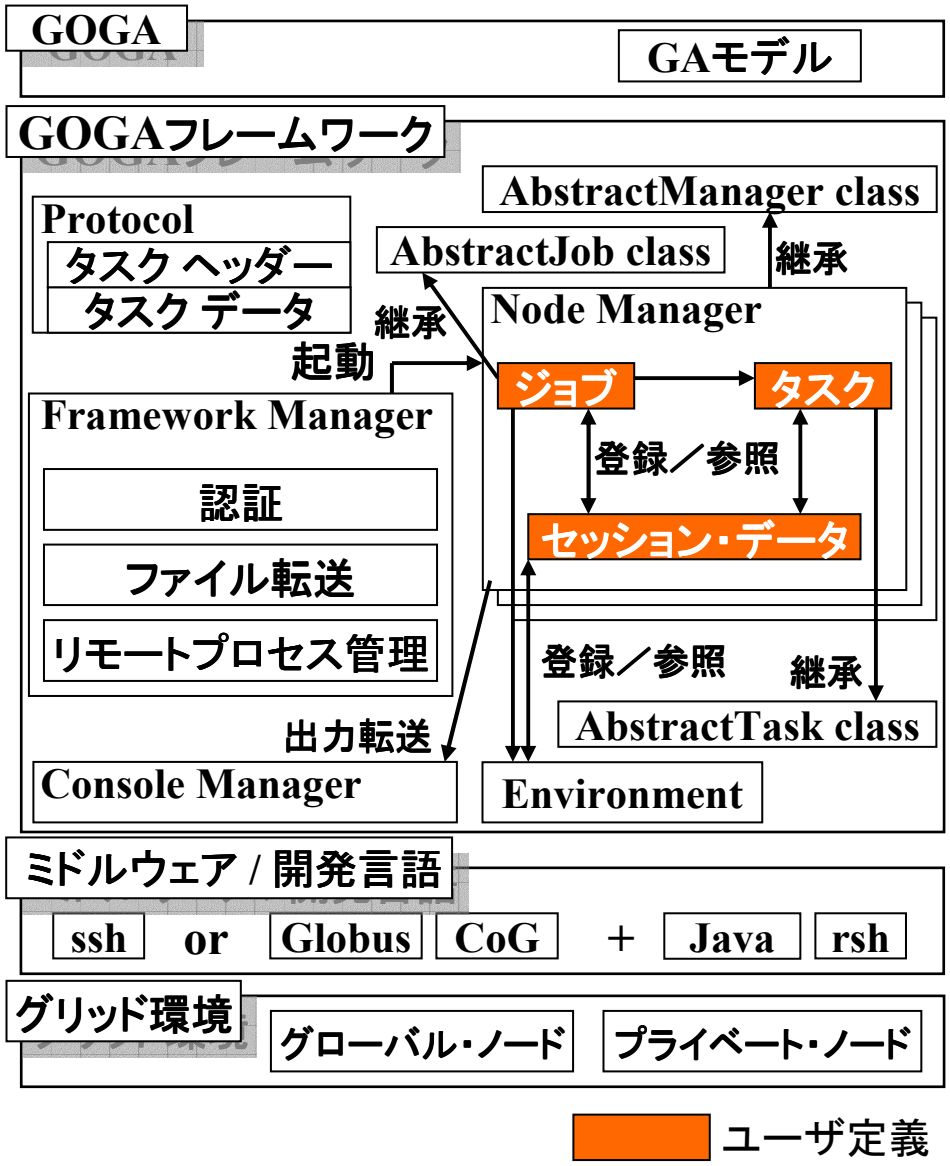
おわりに（2）

専門家向けグリッドポータルシステム



おわりに(3)

グリッド向けGAフレームワーク [水口 05]



謝　辞

- 科学技術振興事業団計算科学技術活用型特定研究開発推進事業「コモディティグリッド技術によるテラスケール大規模数理最適化」(2001～2003年度；東京工業大学 松岡聡教授)
- 東京工業大学の田中氏をはじめとする松岡研究室の皆様
- 東京工業大学の合田助教授をはじめとする合田研究室の皆様
- 東京電機大学の藤沢助教授をはじめとする藤沢研究室の皆様
- 産業技術総合研究所の田中氏をはじめとするグリッドセンターの皆様
- 東京工業大学の小島教授
- 徳島大学の松原君，中井君，多畑君，北村君，林君