

生成 AI を用いた HMMER の「PEZY-SC3」及び「汎用 CPU」向け高速化実装に成功 ～ タンパク質機能解析を支える hmmscan で約 70 倍の高速化を達成 ～

株式会社 PEZY Computing (以下、PEZY Computing) と株式会社 QuettaAI (以下、QuettaAI) は、バイオインフォマティクス分野で広く利用されているシーケンス検索ツール「HMMER[1,2]」を、PEZY Computing が開発・製造した HPC 向け MIMD プロセッサ「PEZY-SC3」及び「汎用 CPU」向けとして高速化実装に成功したことをお知らせいたします。

本プロジェクトはソフトウェアエンジニアである小田 俊之 (QuettaAI 所属) が担当し、最新の生成 AI Coding Agent (Claude[3] Code / Opus 5) を全面的に活用することで、複雑なアーキテクチャへのコード改変を迅速に実現しました。

プロジェクトの背景と AI の活用

昨今の生成 AI の飛躍的な性能向上により、高度な専門知識を要するハードウェア向けのコード改変や最適化を AI が強力にサポートできるようになりました。本取り組みでは、Claude Code に実装プロセスを委ね、HMMER の複雑なコードベースを PEZY-SC3 向けに短期間で最適化することに成功しました。

科学技術計算における信頼性を担保するため、計算結果がオリジナル版と 100% 一致することを条件として開発を行いました。浮動小数点演算の計算順序が問題となる部分を避け、E-value、bitscore など数値部分についても最終出力では厳密に一致することを確認しています。

圧倒的なパフォーマンス向上

各種データセットを用いたベンチマークテストにおいて、劇的な処理速度の向上を確認しました。

- **最大約 70 倍の高速化:** hmmscan 処理において、オリジナル版と比較して最大約 70 倍の高速化を達成しました (ハイスループットシーケンサー (HTS) 由来リード ERR056999[4] を、prodigal[5] により予測した断片 5000 配列処理時、表 1-1)。
- **実用的な解析時間の大幅短縮:** PEZY-SC3 を 4 台搭載した環境を使用することで、1 HTS ラン (ERR056999 を prodigal により翻訳した全リード 262,574 配列) の Pfam[6] アノテーションが約 11 分、1 細菌プロテオーム (大腸菌プロテオーム、4,403 配列) が約 27 秒 (およそ 40 倍の高速化) という圧倒的な速度で処理可能です。hmmsearch を使用する反転テクニック[7]を使用した場合と比較しても約 10 倍の高速化となっています (表 2-1)。
- **CPU 使用でも大幅な高速化:** ホスト側コードの独自最適化効果により、PEZY-SC3 を使用しない CPU 環境下においても、hmmscan で数倍～数十倍の高速化を実現しています。

今後の展開

本プロジェクトで開発された最適化済みソースコードは、オープンソースとして一般公開を予定しております。PEZY-SC3 を含むアクセラレーターを効率的に使用するには高度な専門知識が必要でしたが、近年の Coding AI エージェントの発展により容易にアクセラレーター用コードを実装することが可能となりました。メタゲノム解析をはじめとする膨大な生命科学データ処理のボトルネックを解消し、生命科学分野の研究開発を大きく加速させることが期待されます。

PEZY Computing と QuettaAI は、今後も最先端のハードウェア技術と人工知能技術を掛け合わせ、ハイパフォーマンスコンピューティング（HPC）の新たな可能性を切り拓いてまいります。

本件発行日： 2026 年 9 月 25 日 株式会社 PEZY Computing

本件に関する問い合わせページ： <https://www.pezy.co.jp/contact/>

実行環境、具体的な処理時間等、詳細につきましては下記に記述します。

実行環境

AMD EPYC 7702 64-Core Processor（コア数 64、スレッド数 128）、1 TB RAM ノード。
HMMER 3.4 は README.md に従い、./configure --prefix <path> / make による標準ビルドを行った。

パラメータ探索

--cpu オプションについて、*Escherichia coli* proteome を用いて、オリジナル版は 0（並列化なし）、1（シングルスレッドだが、並列化用のブロックが使用される）、2、4、8、16、32、64、128 を確認した。ただし、CPU 部分改変版、SC3 使用版については、0、1、2 では改善が見込めなかったため 4、8、16、32、64、128 を確認した。

オリジナル版 hmmscan については、--cpu 0、CPU 部分改変版および SC3 使用版 hmmscan については--cpu 64、オリジナル版 hmmsearch については --cpu 4 が最も処理時間が短かった。

Data	配列数	アミノ酸数	Tool	オリジナル版 (--cpu 0)	cpu_ex ^a (--cpu 64)	SC3 (--cpu 64)	CPU高速化率	SC3高速化率
ERR056999 ^b	262574	32571815	hmmscan(Pfam-A 38.2)	NA ^c	1598.306	673.983	NA	NA
ERR056999_5000 ^d	5000	610556	hmmscan(Pfam-A 38.2)	979.648	33.710	14.555	29.061	67.307
hsa_2500 ^d	2500	1375109	hmmscan(Pfam-A 38.2)	831.528	38.823	32.285	21.418	25.756
eco_proteome	4403	1354442	hmmscan(Pfam-A 38.2)	1057.031	41.100	27.318	25.719	38.694
eco100 ^d	100	31686	hmmscan(Pfam-A 38.2)	29.018	3.215	2.933	9.026	9.894
eco100	100	31686	jackhmmer(sprot) ^{e,f}	723.888	173.329	42.403	4.176	17.072
eco100 ^g	100	31686	hmmsearch(sprot) ^{e,f}	1677.944	491.041	113.733	3.417	14.753

表 1-1. 各データセットに対する処理時間。単位は秒。

- a. SC3 を使用せず、CPU 部分の改変だけが適用されたバージョン。
- b. prodigal によるリードからの予測。
- c. 10 時間以上の処理時間が想定されるため割愛。
- d. ERR056999、UniProt[8] よりダウンロードした *Homo sapiens* proteome (UP000005640)、および *Escherichia coli* proteome (UP000000625) からそれぞれ 5000 配列、2500 配列、100 配列をランダムで取ったもの。
- e. swissprot[8] に対する検索。trEMBL[8] など大きい DB が対象である場合は、SC3 利用時でも CPU に Fallback され、CPU 部分改変版とほぼ同じ処理時間となった(結果は割愛する)。
- f. jackhmmer および hmmsearch(sprot) についてはスレッド並列が有効であり、--cpu 32 での比較(表 1-2) ではオリジナル版と CPU 部分改変版はほぼ同等(高速化率約 1.0) となる。
- g. 上の行の jackhmmer による HMM を使用した。

Data	Tool	オリジナル版 (--cpu 32)	cpu_ex ^a (--cpu 32)	SC3 (--cpu 32)	CPU高速化率	SC3高速化率
ERR056999 ^b	hmmScan(Pfam-A 38.2)	NA ^c	1595.11	665.149	NA	NA
ERR056999_5000 ^d	hmmScan(Pfam-A 38.2)	2246.959	34.014	14.719	66.060	152.657
hsa_2500 ^d	hmmScan(Pfam-A 38.2)	1176.821	39.174	32.423	30.041	36.296
eco_proteome	hmmScan(Pfam-A 38.2)	1971.131	41.407	27.607	47.604	71.400
eco100 ^d	hmmScan(Pfam-A 38.2)	45.372	3.266	2.95	13.892	15.380
eco100	jackhmmer(sprot) ^e	475.765	480.393	120.524	0.990	3.947
eco100 ^g	hmmsearch(sprot) ^e	168.961	170.961	45.997	0.988	3.673

表 1-2. 表 1-1 の比較について、--cpu オプションに 32 を指定した場合の結果。配列数・アミノ酸数、脚注は表 1-1 と同じ。

Data	hmmsearch (--cpu 4)	CPU(--cpu 64) 高速化率	SC3(--cpu 64) 高速化率
ERR056999	7816.99	4.891	11.598
ERR056999_5000	174.179	5.167	11.967
hsa_2500	366.68	9.445	11.358
eco_proteome	303.274	7.379	11.102

表 2-1. Pfam アノテーションに hmmsearch を使う反転テクニック[7]を使用した際の、オリジナル版 hmmsearch の処理時間(単位は秒)と、それに対するカスタム版 hmmScan の高速化率。反転テクニックの hmmsearch については--cpu 32 の結果も表 2-2 に併記した。

Data	hmmsearch (--cpu 32)	CPU(--cpu 32) 高速化率	SC3(--cpu 32) 高速化率
ERR056999	9488.088	5.948	14.265
ERR056999_5000	248.975	7.320	16.915
hsa_2500	435.164	11.108	13.421
eco_proteome	394.97	9.539	14.307

表 2-2. 表 2-1 の比較について--cpu オプションに 32 を指定した場合の結果。

- [1] <http://hmmer.org/> (2026/09/24 閲覧)
- [2] Eddy, Sean R. "Accelerated profile HMM searches." *PLoS computational biology* 7.10 (2011): e1002195.
- [3] <https://clade.ai/> (2026/09/24 閲覧)
- [4] https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=ERR056999 を 2026-08-31 に sratoolkit [9]により取得
- [5] Hyatt, Doug, et al. "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC bioinformatics* 11.1 (2010): 119.
- [6] Paysan-Lafosse, Typhaine, et al. "The Pfam protein families database: embracing AI/ML." *Nucleic acids research* 53.D1 (2025): D523-D534.
- [7] <http://cryptogenomicon.org/hmmscan-vs-hmmsearch-speed-the-numerology.html>
(2026/09/24 閲覧)
- [8] Bateman, Alex, et al. "UniProt: the universal protein knowledgebase in 2025." *Nucleic acids research* 53.D1 (2025).
- [9] <https://github.com/ncbi/sra-tools version 3.4.1> (2026/09/24 閲覧)