

HELIOS: High-Speed Sequence Alignment in Optics: S1 Text

EHSAN MALEKI¹, SAEEDAH AKBARI ROKN ABADI¹, AND SOMAYYEH KOOHI^{1,*}

¹Department of Computer Engineering, Sharif University of Technology, Azadi Ave., Tehran, Iran.

* Corresponding author: koohi@sharif.edu

Compiled October 20, 2022

In response to the imperfections of current sequence alignment methods, originated from the inherent serialism within their corresponding electrical systems, a few optical approaches for biological data comparison have been proposed recently. However, due to their low performance, raised from their inefficient coding scheme, this paper presents a novel all-optical high-throughput method for aligning DNA, RNA, and protein sequences, named HELIOS. The HELIOS method employs highly sophisticated operations to locate character matches, single or multiple mutations, and single or multiple indels within various biological sequences. On the other hand, the HELIOS optical architecture exploits high-speed processing and operational parallelism in optics, by adopting wavelength and polarization of optical beams. For evaluation, the functionality and accuracy of the HELIOS method are approved through behavioral and optical simulation studies, while its complexity and performance are estimated through analytical computation. The accuracy evaluations indicate that the HELIOS method achieves a precise pairwise alignment of two sequences, highly similar to those of Smith-Waterman, Needleman-Wunsch, BLAST, MUSCLE, ClustalW, ClustalΩ, T-Coffee, Kalign, and MAFFT. According to our performance evaluations, the HELIOS optical architecture outperforms all alternative electrical and optical algorithms in terms of processing time and memory requirement, relying on its highly sophisticated method and optical architecture. Moreover, the employed compact coding scheme highly escalates the number of input characters, and hence, it offers reduced time and space complexities, compared to the electrical and optical alternatives. It makes the HELIOS method and optical architecture highly applicable for biomedical applications.

1. ACCURACY EVALUATION

In order to comprehensively assess the accuracy of the HELIOS method, two statistical analyses are performed through simulation studies: 1) Quantitative measurement of homology [1], and 2) Accuracy measurement of classification output [2], compared to the well-known algorithms, briefly reported in Tables A1 and A2, respectively. In this study, the *Nine ND5 protein sequences dataset* [3] is assumed in this study, as represented in Table A3.

A. Quantitative measurement of homology

To perform quantitative measurement of homology [1], the parameters Identity, Similarity, and Alignment Score of the HELIOS outputs are calculated through simulation studies, as reported in Tables A4-A6, respectively, assuming the *Nine ND5 protein sequences dataset* [3]. While the Identity reports the number of exactly matched characters of two sequences (in percentage), the Similarity measures the resemblance of two compared sequences. Specifically, regarding the physicochemical properties, the amino acids are categorized into six groups with different Similarity values; including GAVLI, FYW, STCM, KRH, DENQ,

and P. As the third metric, the BLOSUM62 [4] substitution scoring matrix [4] is adopted to calculate the Alignment Score, with gap opening and extension penalties equal to -10 and -0.5, respectively.

For a comparative study, the values of Identity, Similarity, and Alignment Score of the quantitative measurement of homology is performed by various well-known algorithms to be compared to the those of HELIOS method, assuming *Nine ND5 protein sequences dataset* [3]. It includes Smith-Waterman (SW) [5] reported in Tables A7-A9, Needleman-Wunsch (NW) [6] reported in Tables A10-A12, BLAST [7] reported in Tables A13-A15, ClustalW [8] reported in Tables A16-A18, Clustal-Omega [9] reported in Tables A19-A21, MUSCLE [9] reported in Tables A22-A24, T-Coffee [10] reported in Tables A25-A27, Kalign [11] reported in Tables A28-A30, and MAFFT [12] reported in Tables A31-A33.

B. Accuracy measurement of classification output

Afterward, the accuracy measurement of the classification output [2] of the HELIOS method is addressed by calculating the val-

Table A1. A brief report of the quantitative measurement of homology of the HELIOS method, compared to nine well-known algorithms, including SW, NW, BLAST, ClustalW, Clustal Omega, Muscle, T-Coffee, Kalign, and MAFFT. The parameters Identity, Similarity, and Alignment score are averaged and reported. The dataset used in this study is the *Nine ND5 protein sequences dataset* [3].

	HELIOS	SW	NW	BLAST	MUSCLE	ClustalW	ClustalΩ	T-Coffee	Kalign	MAFFT
Identity	76.959	76.580	76.580	76.274	76.292	76.292	76.263	76.344	76.337	76.388
Similarity	83.365	83.623	83.605	83.358	83.325	83.295	83.306	83.358	83.373	83.424
Alignment Score	2314.5	2380.3	2389.9	2373.5	2381.3	2381.0	2380.3	2380.6	2382.3	2383.6

Table A2. A brief report of the accuracy measurement of classification output of the HELIOS method with referencing well-known algorithms, including SW, NW, BLAST, ClustalW, Clustal Omega, Muscle, T-Coffee, Kalign, and MAFFT. The parameters SEN, Spec, Acc, PPV, NPV, MCC, and F-Score are averaged and reported. The dataset used in this study is the *Nine ND5 protein sequences dataset* [3].

	SW	NW	BLAST	MUSCLE	ClustalW	ClustalΩ	T-Coffee	Kalign	MAFFT
SEN	0.94972	0.79205	0.95190	0.95424	0.95491	0.95476	0.95557	0.95561	0.95613
Spec	0.99994	0.99968	0.99994	0.99995	0.99995	0.99995	0.99995	0.99995	0.99990
ACC	0.99985	0.99933	0.99986	0.99987	0.99987	0.99987	0.99987	0.99988	0.99988
PPV	0.96055	0.80181	0.96273	0.96636	0.96704	0.96689	0.96771	0.96775	0.96828
NPV	0.99992	0.99966	0.99992	0.99992	0.99993	0.99993	0.99993	0.99993	0.99993
MCC	0.95503	0.79657	0.95722	0.96020	0.96087	0.96072	0.96155	0.96159	0.96211
F-Score	0.95508	0.79687	0.95726	0.96023	0.96090	0.96075	0.96157	0.96161	0.96210

ues of Sensitivity (SEN), Specificity (Spec), Accuracy (ACC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), Matthew's Coefficient Correlation (MCC), and Test's Accuracy (F-Score) in the simulation studies, according to Eq 5 to Eq 11, respectively.

As a comparative study, the accuracy measurement of the classification output of the HELIOS method is accomplished, assuming *Nine ND5 protein sequences dataset* [3], and the corresponding metrics are calculated with considering Smith-Waterman [5] reported in Tables A34-A40, Needleman-Wunsch [6] reported in Tables A41-A47, ClustalW [8] reported in Tables A55-A61, Clustal-Omega [9] reported in Tables A62-A68, BLAST [7] reported in Tables A48-A54, MUSCLE [13] reported in Tables A69-A75, T-Coffee [10] reported in Tables A76-A82, Kalign [11] reported in Tables A83-A89, and MAFFT [12] reported in Tables A90-A96.

REFERENCES

1. D. S. Moss, S. Jelaska, and S. Pongor, *Essays in bioinformatics*, vol. 368 (IOS Press, 2005).
2. M. Hamada, H. Kiryu, W. Iwasaki, and K. Asai, "Generalized centroid estimators in bioinformatics," *PloS one* **6**, e16450 (2011).
3. M. M. Abo-Elkhier, M. A. Abd Elwahaab, and M. I. Abo El Maaty, "Measuring similarity among protein sequences using a new descriptor," *BioMed research international* **2019** (2019).
4. D. W. Mount, "Using blosum in sequence alignments," *Cold Spring Harb. Protoc.* **2008**, pdb.top39 (2008).
5. H. Zou, S. Tang, C. Yu, H. Fu, Y. Li, and W. Tang, "asw: accelerating smith-waterman algorithm on coupled cpu-gpu architecture," *Int. J. Parallel Program.* **47**, 388–402 (2019).
6. Y. Jararweh, M. Al-Ayyoub, M. Fakirah, L. Alawneh, and B. B. Gupta, "Improving the performance of the needleman-wunsch algorithm using parallelization and vectorization techniques," *Multimed. Tools Appl.* **78**, 3961–3977 (2019).
7. G. M. Boratyn, J. Thierry-Mieg, D. Thierry-Mieg, B. Busby, and T. L. Madden, "Magic-blast, an accurate rna-seq aligner for long and short reads," *BMC bioinformatics* **20**, 1–19 (2019).
8. D. Díaz, F. J. Esteban, P. Hernández, J. A. Caballero, A. Guevara, G. Dorado, and S. Gálvez, "Mc64-clustalwp2: A highly-parallel hybrid strategy to align multiple sequences in many-core architectures," *PLOS ONE* **9**, 1–12 (2014).
9. F. Sievers and D. G. Higgins, "Clustal omega for making accurate alignments of many protein sequences," *Protein Sci.* **27**, 135–145 (2018).
10. C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *J. Mol. Biol.* **302**, 205–217 (2000).
11. T. Lassmann, "Kalign 3: multiple sequence alignment of large datasets," (2020).
12. J. Rozewicki, S. Li, K. M. Amada, D. M. Standley, and K. Katoh, "Mafft-dash: integrated protein sequence and structural alignment," *Nucleic acids research* **47**, W5–W10 (2019).
13. R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.* **32**, 1792–1797 (2004).

Table A3. The list of input sequences, assuming the *Nine ND5 protein sequences dataset* [3].

Name	Sequence
Homo sapiens	MTMHT TMTTL TLTSI IPPIL TTLVN PNKKN SYPHY VKSIV ASTFI ISLFP TTMFM CLDQE VIISN WHWAT TQTTQ LSLSF KLDYF SMMFI PVALF VTWSI MEFSL WYMNS DPNIN QFFKY LLIFL ITMLI LVTAN NLFQL FIGWE GVGIM SFLLI SWWYA RADAN TAAIQ AILYN RIGDI GFILA LAWFL LHSNS WDPQQ MALLN ANPSL TPLLG LLAA AGKSA QLGLH PWLPS AMEGP TPVSA LLHSS TMVVA GIFLL IRFHP LAENS PLIQT LTLCL GAITT LFAAV CALTQ NDIKK IVAFS TSSQL GLMMV TIGIN QPHLA FLHIC THAFF KAMLF MCSGS IIHNL NNEQD IRKMG GLLKT MPLTS TSLTI GSLAL AGMPF LTGFY SKDHI IETAN MSYTN AWALS ITLIA TSLTS AYSTR MILLT LTGQP RFPTL TNINE NNPTL LNPIK RLAAG SLFAG FLITN NISPA SPFQT TIPLY LKLLA LAVTF LGLLT ALDLN YLTNK LKMKX PLCTF YFSNM LGFYP SITHR TIPYL GLLTS QNLPL LLLDL TWLEK LLPKT ISQHQ ISTSI ITSTQ KGMK LYFLS FFFPL ILTLL LIT
Rattus norvegicus	MNMMT ILILM ILDDL TTPIT FSMIT LTKLM YFRHM LITSI KFSFL LSLLP LLLFF HHNTE YMITN WHWLT INSIK LTMSF KIDYF SILFL SVSLF VTWSI MQFSS WYMHS DPHIN RFIKY LMMFL NNMLI LTSAN NLFQL FIGWE GVGIM SFLLI GWWYG LADAN TAAIQ AILYN RVGDI GFILA MTWFC LNMNS WELQQ IFLTN TNNNL VPLTG LLIAA TGKSA QFGLH PWLPS AMEGP TPVSA LLHSS TMVVA GIFLM IRFHP LTSNN STIMT AMLCL GAITT LFTAI CALTQ NDIKK IVAFS TSSQL GLMMV TLGIN QPYLA FLHIC THAFF KAMLF MCSGS IIHIL NDEQD IRKMG NMMKA MPPTS SCLII GSLAL TGMPF LTGFY SKDLI IEAIN TCNTN AWALM ITLIA TSMTA VYSMR IYFV TMTKP RYSPL ITINE NNPPL INPIK RLALG SILAG FLISL NIPPT NIQIL TMPWH LKMTA LLITI LGFAI ALELN NLTLN LSMK PTKLS SFSTS LGYYP PIMPR IIPQK TLNSS YKLSL NLLDL IWSEK TIPKS TSITQ TQLSK MMSNQ KGLIK LYFLS FLITI SLIFI LHTLN PEWFQ
Balaenoptera physalus	MNLFT SFTLL TLLIL TTPIM MSHTG SHVNN KYQSY VKNIV FCAFI TSLVP AMVYL HTNQE TLISN WHWIT IQTLK LTLSF KMDYF SLMFM PVALF ITWSI MEFSM WYMHS DPNIN QFFKY LLLFL ITMLI LVTAN NLFQL FIGWE GVGIM SFLLI GWWFG RTDAN TAAIQ AILYN RIGDI GLLAS MAWFL SNMNT WDLQ IFMLN QNPLN FPLMG LVLA AGKSA QFGLH PWLPS AMEGP TPVSA LLHSS TMVVA GIFLL VRFYP LMENN KLIQT VTLCCL GAITT LFTAI CALTQ NDIKK IAFS TSSQL GLMMV TIGLN QPYLA FLHIC THAFF KAMLF LCSGS IIHNL NNEQD IRKMG GLFKA LPFTT TALII GCLAL TGMPF LTGFY SKDPI IEAAT SSYTN AWALL LTLTA TSLTA VYSTR IFFA LLGQP RFPFS TTINE NNPL INPIK RLLVG SIFAG FILSN SIPP MTPML TMLPH LKLLA LAMTT LGFII AFEIN LDTQN LKHKH PSNSF KFSTL LGYYP TIMHR LPPHL DLLMS QKLAT SLLDL TWLET ILPKT TALIQ LKAST LTSNQ QGLIK LYFLS FLITI TLSMI LFNYPE
Balaenoptera musculus	MNLFT SFVLL TLLIL FTPIM VSNTD PHKNN KYQSY VKNIV FCAFI TSLIP AMMYL HTNQE TLISN WHWIT IQTLK LTLSF KMDYF SLMFM PVALF ITWSI MEFSM WYMHS DPNIN QFFKY LLLFL ITMLI LVTAN NLFQL FIGWE GVGIM SFLLI GWWFG RTDAN TAAIQ AILYN RIGDI GLLAS MAWFL SNMNT WDLQ IFMLN QNPLN FPLMG LVLA AGKSA QFGLH PWLPS AMEGP TPVSA LLHSS TMVVA GIFLL VRFYP LMENN KLIQT VTLCCL GAITT LFTAI CALTQ NDIKK IAFS TSSQL GLMMV TIGLN QPYLA FLHIC THAFF KAMLF LCSGS IIHNL NNEQD IRKMG GLFKA LPFTT TALII GCLAL TGMPF LTGFY SKDPI IEAAT SSYTN AWALL LTLTA TSLTA VYSTR IFFA LLGQP RFPFS TTINE NNPL INPIK RLLIG SIFAG FILSN SIPPV ITPLM TMLPH LKLLA LAMTT LGFII AFEIN LDTQN LKTH PSNPF KFSTL LGYYP TIMHR LPPHL DLSMS QKLAT SLLDL TWLET ILPKT TALIQ LKAST LTSNQ QGLIK LYFLS FLITI TLSMI LFNCPE
Didelphis virginiana	MKVIN ISNTM SIMSI ILLIL PLLYN LSLK KINFP LYCKN MIMLA FMMSL PSLLL FMYKG QESII TNWHW FSIS FNISM SFKMD FFSII FIPIA LFVTW AILEF SLWYM HSDPN ISQFF KYLII FLLTM IILVS ANNLF QLFIG WEGVG IMSFL LIGWW YGRSD ANTAA LQAIL YNRIG DIGFM LTMAY LMLNC NSWDL QHIFS MNMHP IALLG LLIAA TGKSA QFSLH PWLPS AMEGP TPVSA LLHSS TMVVA GIFLL IRFHP MLENN KTMILT ITLCL GALT LFTAM CAIMQ NDIKK IVAFS TSSQL GLMMV TVGLN QPHLA FLHIC THAFF KAMLF LCSGS IIHNL NNEQD IRKMG GLFYT LPITS SALMT GSLAL MGTPF LAGFY SKDSI IEAMN TSYTN SWALT ITLIA TSLTA IYSLR IYYT LLGHP RFMTM SPLNE NNPPL INPI RLALG TIFAG FMLTT NMPPS YSITM TMPMF IKQMA LMVTT TGLMM GMELN SLTNK LMKSN NHTNN FLTML GFYTQ IMHRM QPLIS LFMGQ RIATM LIDMN WYEKA GPKGQ ANVHS KLSSL ISSSQ KGLMK MYFLS FLVSM IFIIL FT
Pan troglodytes	MTMYA TMTTL ALTSL IPPIL GALIN PNKKN SYPHY VKSII ASTFI ISLFP TTMFM CLDQE TIISN WHWAT TQTTQ LSLSF KLDYF SMTFI PVALF VTWSI MEFSL WYMNS DPNIN QFFKY LLIFL ITMLI LVTAN NLFQL FIGWE GVGIM SFLLI SWWYA RTDAN TAAIQ AILYN RIGDI GFILA LAWFL LHSNS WDPQQ MILLS TNTDL TPLLG FLAA AGKSA QLGLH PWLPS AMEGP TPVSA LLHSS TMVVA GIFLL IRFYP LAENN PLIQT LTLCL GAITT LFAAV CALTQ NDIKK IVAFS TSSQL GLMMV TIGIN QPHLA FLHIC THAFF KAMLF MCSGS IIHNL NNEQD IRKMG GLLKT MPLTS TSLTI GSLAL AGMPF LTGFY SKDLI IETAN MSYTN AWALS ITLIA TSLTS AYSTR MILLT LTGQP RFPTL TNINE NNPTL LNPIK RLTI G SLFAG FLITN NILPM STPQV TIPLY LKLLA LGVTS LGLLT ALDLN YLTSK LKMKX PLYTF HFSNM LGFYP NIMHR SIPYL GLLTS QNLPL LLLDL TWLEK LLPKT ISQYQ ISASI TTSTQ KGMK LYFLS FFFPL ILTLL LIT
Pan paniscus	MTMYT TMTTL TLTPL ILPII TTLIN PNKKN SYPHY VKSII ASTFI ISLFP TTMFM CLDQE AIISN WHWAT TQTTQ LSLSF KLDYF SMTFI PVALF VTWAI MEFSL WYMNS DPNIN QFFKY LLIFL ITMLI LVTAN NLFQL FIGWE GVGIM SFLLI SWWYA RTDAN TAAIQ AILYN RIGDI GFILA LAWFL LHSNS WDPQQ MVLLS TNPSL TPLLG FLAA AGKSA QLGLH PWLPS AMEGP TPVSA LLHSS TMVVA GVFLI IRFHP LAENN PLIQT LTLCL GAITT LFAAI CALTQ NDIKK IVAFS TSSQL GLMMV TIGIN QPHLA FLHIC THAFF KAMLF MCSGS IIHNL NNEQD IRKMG GLLKT MPLTS TSLII GSLAL AGMPF LTGFY SKDLI IETAN MSYMN AWALS ITLIA TSLTS AYSTR MILLT LMGQP RFPTL TNINE NNPTL LNPIK RLTI G SLFAG FFITN NILPM STSQM TIPLY LKLLA LSITL LGLLT ALDLN YLTNK LKMKX PPYTF YFSNM LGFYP NIMHR SIPYL GLLTS QNLPL LLLDL TWLEK LLPKT ISQYQ VSASI TTSTQ KGMK LYFLS FLPL ILTLL LIM
Gorilla gorilla	MTMYA TMTTL ALTSL IPPIL TTFIN PNKKS SYPHY VKSIV ASTFI ISLFP TTMFL CLDQE AIISN WHWAT TQTIQ LSLSF KLDYF SMMFI PVALF VTWSI MEFSL WYMNS DPNIN QFFKY LLIFL ITMLI LVTAN NLFQL FIGWE GVGIM SFLLI GWWYA RTDAN TAAVQ AILYN RIGDI GFILA LAWFL LHSNS WDPQQ MSLN TNPNL IPLLG FLAA AGKSA QLGLH PWLPS AMEGP TPVSA LLHSS TMVVA GVFLI IRFRH LAENN SLAQT LTLCL GAITT LFAAV CALTQ NDIKK IVAFS TSSQL GLMVA TIGIG QPHLA FLHIC THAFF KAMLF MCSGS IIHNL NNEQD IRKMG GLLKA MPLTS TSLAI GSLAL MGMPF LTGFY SKDLI IETAN MSHTN AWALS IILIA TSLTS AYSTR MILLT LTGQP RFPTF ANINE NYSTL LNPIK RLTI G SLFAG FFITN NILPT SVPQM TIPLY LKLLA LSITL LGLLT ALDLN YLTNK LKMKH PHTF YFSNM LGFYP NITHR TIPYL GLLMS QNLPL LLLDL IWLEK LLPKT ISQHQ ISASI TTSTQ KGLIK LYFLS FFFPL LLILL LIT
Mus musculus	MNIPT TSILL IFILL LSPIL ISMSN LIKHI NFPLY TTTSI KFSFI ISLLP LLMFF HNNME YMITT WHWVT MNSME LKMSF KTDFF SILT SVALF VTWSI MQFSS WYMNS DPNIN RFIKY LTLFL ITMLI LTSAN NMFQL FIGWE GVGIM SFLLI GWWYG RTDAN TAAIQ AILYN RIGDI GFILA MVWFS LNMNS WELQQ IMFSN NNDNL IPLMG LLIAA TGKSA QFGLH PWLPS AMEGP TPVSA LLHSS TMVVA GIFLL VRFHP LTTNN NFILT TMLCL GALT LFTAI CALTQ NDIKK IAFS TSSQL GLMMV TLGMN QPHLA FLHIC THAFF KAMLF MCSGS IIHSL ADEQD IRKMG NITKI MPPTS SCLVI GSLAL TGMPF LTGFY SKDLI IEAIN TCNTN AWALL ITLIA TSMTA MYSMR IYFV TMTKP RFPTL ISINE NDPDL LNPIK RLAFG SIFAG FVSY NIPPT SIPVL TMPWF LKTTA LIISV LGFLI ALELN NLTMK LSMNK ANPYS SFSTL LGFFP SIHR ITPMK SLNLS LKTSI LLLDL IWLEK TIPKS TSTLH TNMTT LTTNQ KGLIK LYFMS FLINI ILIII LYSIN LE

Table A16. The parameter Identity of the ClustalW in the quantitative measurement of homology, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A17. The parameter Similarity of the ClustalW in the quantitative measurement of homology, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A18. The parameter Alignment Score of the ClustalW in the quantitative measurement of homology, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A28. The parameter Identity of the Kalign in the quantitative measurement of homology, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A29. The parameter Similarity of the Kalign in the quantitative measurement of homology, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A30. The parameter Alignment Score of the Kalign in the quantitative measurement of homology, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A61. The parameter Test's Accuracy (F-Score) of the HELIOS method with referencing the ClustalW in the accuracy measurement of classification output, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A62. The parameter Sensitivity (SEN) of the HELIOS method with referencing the ClustalΩ in the accuracy measurement of classification output, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

Table A63. The parameter Specification (Spec) of the HELIOS method with referencing the ClustalΩ in the accuracy measurement of classification output, assuming the *Nine ND5 protein sequences dataset* [3].

[illegible]

