**PROPOSED TOPICS FOR MASTER THESIS**

Program: Systems biology

Admission year: 2020

Published: October, 2020

|  |  |  |
| --- | --- | --- |
| **Topic with description** | **Contact person** | **Contact email** |
| **TITLE: Computational mapping of nucleosome positions across variety of organisms**  ABSTRACT. Nucleosome positioning DNA sequence patterns (NPS) - usually distributions of particular dinucleotides or other sequence elements in nucleosomal DNA - at least partially determine chromatin structure and arrangements of nucleosomes that in turn affect gene expression. Statistically, NPS are defined as oscillations of the dinucleotide periodicity with about 10 base pairs (bp) which reflects the double helix period. Recently few distinctive patterns in nucleosomal sequences were observed that can be termed as packing and regulatory referring to distinctive modes of chromatin function [1]. Our working hypothesis for the future studies is that packing patterns tend to be preferred by evolutionary lower organisms and regulatory ñ by higher organisms.  Given vast amount of publicly available nucleosome maps in various organisms [2,3] it is possible to perform a computational mapping of nucleosomes at the loci of gene promoters in various organisms using already available dnpatterntools v1.0 software (https://github.com/erinijapranckeviciene/dnpatterntools).  Although such mapping was attempted previously, the results were inconclusive because of a small number of available nucleosomal DNA seqences at that time [4].  Results of a new computational experiment using data from NucMap [4] and NucPosDB[3] would provide more iformation on pattern preferences across organisms and subsequently better prediction of nucleosome's position at a specific loci. This would eventually lead to the better tools of computational mapping of nucleosome positions.  REFERENCE  1. Nucleosome positioning sequence patterns as packing or regulatory. Erinija Pranckeviciene, Sergey Hosid, Nathan Liang, Ilya Ioshikhes. bioRxiv 755272;doi:https://doi.org/10.1101/755272  [Accepted for publication in PLOS Computational Biology]  2. Teif V.B. (2016). Nucleosome positioning: resources and tools online. Briefings in Bioinformatics 17, 745-757. [https://generegulation.org/nucleosome-positioning-database/]  3. Yongbing Zhao, Jinyue Wang, Fang Liang, Yanxia Liu, Qi Wang, Hao Zhang, Meiye Jiang, Zhewen Zhang, Wenming Zhao, Yiming Bao, Zhang Zhang, Jiayan Wu, Yan W Asmann, Rujiao Li, Jingfa Xiao, NucMap: a database of genome-wide nucleosome positioning map across species, Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D163ñD169, https://doi.org/10.1093/nar/gky980  [https://bigd.big.ac.cn/nucmap/]  4. Liang N, Pranckevicience E and Ioshikhes I. Matching nucleosome patterns to the complexity of the organism [version 1; not peer reviewed]. F1000Research 2019, 8:127 (poster) (https://doi.org/10.7490/f1000research.1116423.1) | Erinija Pranckevičienė | erinija.pranckeviciene@mf.vu.lt |
| **TITLE: Molecular shape similarity in virtual screening for drug discovery**  ABSTRACT. Virtual Screening can drastically accelerate drug discovery processes. Molecular shape similarity is essential in virtual screening for drug discovery. Shape similarity is used to compare in detail the shape of a query molecule against a large database of potential drug compounds. In order to evaluate shape similarity accurately the molecules should be optimally adjusted. In this work, optimization problems for molecular shape similarity are investigated aiming at fast solution and acceptable accuracy. | Julius Žilinskas | julius.zilinskas@mif.vu.lt |
| **TITLE: Comprehensive Pathology Analytics of Renal Tissue Based on Segmentation of Multi-stained Tissue Sections.**  ABSTRACT. The aim of this study is to develop a microscopy image analytics platform for kidney pathology quantification. At the core of the platform is the segmentation of the kidney morphology from tissue sections subjected to different immunohistochemical stainings, each stain optimizing the delineation of a different kidney morphology. Whole slide kidney microscopy images with annotations of different tissue components are available for both the machine/deep learning segmentation and the development of a comprehensive set of quantitative kidney pathology indicators. We expect the use of Python for the machine learning part and R as the statistical back-end of the analytics pipeline, and we look forward to a motivated and focused applicant to join our digital pathology team. | Arvydas Laurinavičius | arvydas.laurinavicius@vpc.lt |
| **TITLE: Patterns of adaptive positive selection in human populations from high-density SNP data.**  ABSTRACT. The analysis of geographically specific regions and the characterization of fine-scale patterns of genetic diversity may facilitate a much better understanding of the microevolutionary processes affecting local human populations.  Geographically specific microevolutionary processes can be inferred when exploring local patterns of population structure and adaptation within the global and historical genetic context established from large general population and ancient publicly available reference panels.  The Lithuanian population is partially isolated with genetic distinctiveness within the European context and with preserved ancient genetic composition. A small genetic differentiation among the Lithuanian ethnolinguistic groups exist (Urnikyte et al., 2019).  Aim of the practice: to analyse signatures of adaptive positive selection in high-density genotyping data previously generated in two main ethnolinguistic population groups from Lithuania. The student will perform the main following tasks:  Task 1. Parsing and transforming data generated in six Lithuanian populations with the Illumina Infinium whole-genome SNP array to appropriate formats for subsequent analyses.  Task 2. Compiling, parsing and transforming external reference population datasets including the CEU (Utah residents with ancestry from northern and western Europe), FIN (Finnish in Finland), TSI (Toscani in Italy) and GBR (British in England and Scotland) populations as wells as additional reference populations from Africa, East Asia and South Asia available from the 1000 Genomes Project (http://www.internationalgenome.org/about/).  Task 3. Performing population structure analysis of six Lithuanian ethnolinguistic groups by PCA using different regional European and worldwide global contexts.  Task 4. Calculating kinship and inbreeding coefficients in the Lithuanian population.  Task 5. Computing selection tests based on population differentiation (XP-EHH statistic, FST ) and site frequency spectrum (Tajima’s D) from SNP genotyping phased/unphased data and using different pairwise population comparisons when appropriate to identify and distinguish signals of positive selection specific of Lithuanians from adaptations shared with other Europeans populations.  Task 6. Identifying top significant candidate genomic windows for positive selection, as wells as pathways enriched for signals of positive selection in the Lithuanian population.  Task 7. Detecting potential functional variation linked to the signatures of selection identified using different functional element annotations of the human genome and diverse in silico prediction methods (ANNOVAR, RefSeqGene, dbSNP147 and CADD).  The student will acquire basic knowledge and skills on the analysis of genetic structure on human populations as well as on the detection of natural selection from high-density genome-wide SNP data. population genetics classic software such as Plink, shapeit, selscan, popgenome and annovar among others.  Requirements: Bioinformatic skills. | Alina Urnikytė | alina.urnikyte@mf.vu.lt |
| **TITLE: Application of Deep Neural Networks for Eye Fundus Image Analysis**  ABSTRACT: Aim of the study is to develop workflow for the eye fundus images analysis with the view to detect various anatomical changes. Early detection of the changes can indicate and even prevent such diseases as diabetes, glaucoma, cataract and etc. The images of known structure changes will be available as well as the consultations with the skilled ophthalmologist. Proposed solution has to be built on deep neural network application and tested. Students selecting the theme will face the Python as a primary programming language that will utilize the Keras and Tensorflow frameworks as those mostly used tools for deep neural network application with the strong community support. The student should have strong motivation to learn new skills all other knowledge will be acquired during the master study. | Povilas Treigys | povilas.treigys@mif.vu.lt |
| **Gardelės parametrų išgavimas iš žmogui skirtų tekstų ir COD apžvalginės DB kūrimas**  **Extracting Unit cell parameters of crystals from human readable scientific papers**  LT: Deja, kol kas ne visuose straipsniuose paskelbtos kristalų struktūros yra atvirai aprašytos viešai prieinamais duomenimis. Daugeliui tokių „paslėptų“ struktūrų straipsniuose nurodomi pagrindiniai kristalo parametrai – gardelės konstantos, simetrijos grupė, molekulės cheminė formulė ir pavadinimas. Tai, nors ir nepilni, bet vis viena labai vertingi duomenys, nes leidžia preliminariai identifikuoti kristalinę medžiagą. Darbe bus pasiūlyta surinkti visų publikuotų kristalų struktūrų bibliografijas ir pagal atvirai prieinamus straipsnių tekstus arba santraukas (angl. abstracts), naudojant paprastas Perl reguliarias išraiškas (o vėliau gal būt ir sudėtingesnę lingvistinę analizę) automatiškai išgauti kiekviename straipsnyje publikuotų medžiagų kristalų parametrus.  Tiriamasis darbo aspektas: darbe bus keliama užduotis atsakyti į klausimus a) kiek ir kokių kristalų struktūrų yra publikuota patikimoje mokslo spaudoje b) kokia jų dalis tinkamai reprezentuota kristalografinėse duomenų bazėse c) kokios informatikos metodikos (algoritmai, sistemos) geriausia tinka gardelės parametrų išgavimui iš spausdintų tekstų d) koks šių automatinių duomenų išgavimo metodų tikslumas ir pilnumas (angl. recall/precision)  EN: Unfortunately, not all research data are publicly openly available for the published articles in a systematised machine readable form, and crystallography is no exception. Some data about solved and published crystal structures are available only behind a paywall. For many of such “hidden” structures, the basic parameters of the crystal are however published in the papers themselves; these parameters include unit cell parameters, temperature and pressure at which these were measured, crystal symmetry space group and chemical composition. These data would be highly valuable if collected in a computer readable database: they would permit identification of a crystalline material, correlation of properties with cell constants and symmetries, and possibly restoration of 3D structure using quantum mechanical computations.  In this work we will develop the system that would scan automatically the available open access literature (papers, preprints, theses) in fields of chemistry, crystallography, physics and material science, extract the afore-mentioned data using various methods (from simple regular expressions to more sophisticated linguistic analysis to trained ANNs) and organise the data in a searchable database. Investigations of different methods human language analysis methods with respect to information recall and accuracy will have to be made, and recommendations as to which methods are most suitable for the task should be derived. The collected data will be used to answer questions regarding what are the most widespread material and crystal types that are investigated, what percentage of them is available with full 3D coordinates, which cell sizes and space groups are prevalent for different material types. | Saulius Gražulis | grazulis@ibt.lt |
| **Informacijos išgavimas iš mokslinių straipsnių**  **Extracting scientific facts from research papers**  LT: Pastaruoju metu publikuojamų mokslinių straipsnių skaičius auga eksponentiškai [1]. Tikėtina, kad šiuo metu straipsnių paskelbiama tiek daug, jog net ekspertai negali fiziškai perskaityti visų straipsnių savo darbų tematika, jau nekalbant apie gretimų tyrimo krypčių straipsnius. Mokslo žinios tampa išsklaidytos po daugelį žmogui sunkiai aprėpiamų šaltinių, jas vis sunkiau susisteminti, o tarpdisciplininiai tyrimai – vienas svarbių inovacijos šaltinių – dėl šių priežasčių neatskleidžia viso savo potencialo. Turėtų būti įmanoma sukurti apmokytus DNT ir/arba kitas kompiuterines sistemas, gebančias iš žmogui skirto teksto (mokslinio straipsnio, patento, monografijos, disertacijos, preprinto) išgauti faktinę mokslinę informaciją, pvz. medžiagos formulę, kristalo ir molekulės struktūros aprašymą, molekulės chemines ir fizines savybes, tų savybių epistemologinį statusą (matavimo rezultatas, teorinis skaičiavimas), ir pateikti šią informacija formalizuotu, automatiniam apdorojimui tinkamu pavidalu, pavyzdžiui, formaliai specifikuotų CIF, CML a SDF failų pavidalu, reliacinių duomenų lentelių pavidalu, RDF failų pavidalų naudojant kurią nors žiniatinklio ontologiją ar pan. Darbe bus pasiūlyta, pasitelkiant įvairias teksto analizės priemones (teksto segmentatorius, simbolių atpažinimo programas, tokias kaip 'tesseract', dirbtinius neuroninius tinklus, mašinų mokymo programas) išgauti iš mokslinių straipsnių vertingus faktus ir apjungti juos į sistemingą duomenų bazę greitai ir išsamiai paieškai. Tiriamasis darbo aspektas: darbe bus keliama užduotis atsakyti į klausimus a) kokie metodai geriausiai tinka mokslo duomenims struktūrizuota (spec. failų formatai, duomenų bazės, etc.), kokie yra įvairių metodų privalumai ir trūkumai b) kokios informatikos metodikos (algoritmai, sistemos) informacijos išgavimui iš spausdintų tekstų ir jos pateikimui struktūrizuota forma, kiekybiškai įvertinant jų privalumus ir trūkumus (tikslumą, pilnumą, greitaveiką)  EN: The amount of currently published scientific papers grows exponentially in all major fields of knowledge (crystallography, medicine, biochemistry, systems biology). The amount of material is larger than people, even experts of the field, are capable of processing and analysing even for their own field, not to mention other adjacent disciplines. As interdisciplinary research becomes more important, such situation hampers progress of new knowledge acquisition. Most importantly, experimental data is often scattered in multiple publications in form of human-oriented tables, graphs and diagrams, precluding systematic data search over the large corpus of data.  In this work we will suggest exploring various methods of extraction of information, most importantly, data values, from papers using various tools: text segmentation, machine learning, OCR, specialised file formats (XML, HTML, CIF, JSON). The questions to be answered are what amount of data can be extracted from publications and what are the best ways to organise it into searchable scientific databases (relational, noSQL, distributed, etc.). Methods of data extraction should be evaluated quantitatively (regarding their recall, precision, speed); their pros and cons should be evaluated. | Saulius Gražulis | [grazulis@ibt.lt](mailto:grazulis@ibt.lt) |
| **Naršyklės įskiepis kristalografinei informacijai rinkti**  **Browser plugin for collection of scientific information**  LT: Šiuolaikiniame mokslo pasaulyje labai didelė informacijos dalis yra išbarstyta po daugybę teksto puslapių (mokslinių straipsnių, disertacijų, monografijų), ir jų sisteminga paieška kompiuterinėmis priemonėmis yra labai apsunkinta. Deja, automatinė tekstų analizė ne visada sugeba teisingai išgauti reikiamus duomenis, ir tenka įdėti daug žmogaus rankų darbo, įvedant informaciją į kompiuterizuotas duomenų bazes. Siūloma sukurti įrankį, kuris palengvintų tyrėjams informacijos surinkimą ir kaupimą vienoje gerai struktūruotoje duomenų bazėje. Siūlomas darbas būtų skirtas įskiepiui, kuris padeda rinkti kristalografinę ir cheminę informaciją, sukurti. Tai leistų šimtams tyrėju vienu metu visame pasaulyje rinkti informaciją, dėti ją į bendrą duomenų bazę ir po to bendrai naudotis surinktais duomenimis. Toks modelis jau neblogai pasiteisino renkant citavimo ir bibliografinius duomenis; naudinga būtų pritaikyti jį kitoms mokslo duomenų rūšims. Tiriamasis darbo aspektas: darbe bus keliama užduotis atsakyti į klausimus: a) kokia yra publikuojamos kristalografinės informacijos apimtis ir kaip tą informaciją panaudoti patikimoms mokslinėms išvadoms gauti, naudojant šiuolaikines kompiuterines technologijas b) kaip paspartinti mokslininko darbą, renkant ir sisteminant reikimą tyrimams informaciją, kokie kompiuterinių technologijų ir naudotojo sąsajos metodai yra tinkamiausi šiam tikslui.  EN: In the current scientific works information is scattered through vast number of publications, often available at many different Web sites, journals, monographs. Since these works are represented as human-dedicated texts in various formats, systematic search of information using computerised means becomes problematic; library catalogues allow searches using keywords or other criteria established by librarians, but do not permit easily to search data using numeric or symbolic properties (e.g. crystal unit cell parameters, electric or magnetic property values) that are important for scientists. Collecting such information into one open searchable database would have enormous value for scientists and general public, allowing quick and accurate access to the latest important scientific research. Automatic collection of such information is not always possible and even when possible, does not always yield accurate results; thus, human help would be needed. Human time, however, is an expensive resource, thus it makes sense to use it as efficiently as possible and to pool and share the results produced by numerous contributors. One method of organising a collaborative network would be to use the World Wide Web and browser plugins that would enable efficient collection, sharing ans searching of information at your fingertips. Such or similar methods were successfully used for citation search (Unpaywall) or in collaborative projects of citizen science like Zooniverse. W propose to explore applicability of this method for collecting data in the field of Systems Biology, Crystalllography, Materials Science and Bioi-/Cheminformatics. A browser plugin should be written or adapted from existing F/LOSS repositories, and different configurations should be explored regarding convenience for human operator, willingness of researchers to participate and share data, ease and convenience of use. In case of successful data collection, question should be investigated what is the most efficient way to represent and share the collected data, what are the value/effort trade-offs for different data items. The collected data then can be applied for automatic searches, training machine learning systems, automatic inference and building interconnected semantic networks with existing data repositories, such as COD, PDB, WikiData, Pubchem and others. | Saulius Gražulis | [grazulis@ibt.lt](mailto:grazulis@ibt.lt) |
| **DNT ir mašinų mokymo pritaikymas kristalų savybėms prognozuoti**  **Applying ANN and machine learning for crystal property prediction**  Nauji Dirbtinių neuroninių tinklų (DNT) ir mašinų mokymo algoritmai leidžia aptikti dėsningumus ir atpažinti bruožus didelėse duomenų masyvuose, kurie seniau buvo neprieinami išsamiai analizei. Atviroje kristalografinėje duomenų bazėje Crystallography Open Database (COD, https://www.crystallography.net/) yra sukaupta virš 450 tūkst įrašų apie kristalų struktūras, o susieti straipsniai talpina informaciją apie šių kristalų savybes. Darbe bus siūloma panaudoti COD DB duomenų imtį DNT ar mašinų mokymo sistemai apmokyti, siekiant prognozuoti įvairias kristalo savybes (pvz. kristalo elementaraus narvelio tūrį, lydymosi temperatūrą ir pan.). Gauti tinklai gali būti naudojami COD ir kitų publikuotų duomenų validavimui, naujų kristalų savybių nustatymui.  EN: Modern methods of machine learning allow to detect latent features and regularities in large amounts of data which were formerly inaccessible for automated analysis. The Crystallography Open Database (COD, <https://www.crystallography.net/>), the world’s largest collection of open access crystal structure data for small molecules. This data set open excellent opportunity to train various machine learning tools (ANNs, SVMs, Bayesian classifiers) to predict various properties of materials, starting from unit cell volume, melting point, dielectric properties from material structures. In this work, several different machine learning methods should be quantitatively investigated for their suitability for the crystal property prediction tasks. The trained networks should then be used to validate the COD data base and incoming structures to detect possible inaccuracies or errors in data. | Saulius Gražulis | [grazulis@ibt.lt](mailto:grazulis@ibt.lt) |
| **Interaktyvios COD recenzavimo svetainės kūrimas**  **Interactive collaboration platform for the Crystallography Open Database**  Interaktyvios programų archyvų svetainės, tokios kaip GitLab, GitGub ar BitBucket, gerai užsirekomendavo programinės įrangos kūrimo procese. Darbe bus siūloma pritaikyti analogiškus programų kūrimo įrankius mokslo duomenims tvarkyti. Atvira kristalografinė duomenų bazė Crystallography Open Database (COD, https://www.crystallography.net/) sėkmingai naudoja Subversijos versijų kontrolės sistemą duomenims versijuoti ir kaupti. Natūralu panaudoti sistemą, analogišką aukščiau minėtoms priemonėms, pvz. Redmine, kuri leistų kristalografams aptarinėti ir taisyti kristalų struktūrų duomenis, panašiai kaip programuotojai aptarinėja ir taiso programas. Darbo metu reikės sukurti COD ir Redmine svetainių sąsają ir išbandyti kristalografinių duomenų valdymo srautą realaus gyvenimo sąlygomis.  EN: The quality of scientific publications is ensured first of all through a peer review process. A similar process is now applied using interactive Web tools for software development (Gitlab, Github, Bitbucket, Redmine). Unfortunately, the review of published scientific data (as opposed to scientific paper text) is often performed very superfluously or not at all. This opens possibilities for cases of scientific fraud with faked data. To minimise such possibilities and ensure highest possible quality of scientific databases, this work suggests implementation of an interactive review system for the Crystallography Open Database (COD, <https://www.crystallography.net/>) repository. Investigations regarding efficiency of problem detection, convenience of use, willingness of scientists to participate. Different models of peer-review (anonymous, open, double-blind) should be explored. | Saulius Gražulis | [grazulis@ibt.lt](mailto:grazulis@ibt.lt) |
| **COD P1 narvelių skaičiavimas**  **Computing chemical structures in P1 cell for the Crystallography Open Database**  Kristalografiniai duomenų failai pateikia minimalų parametrų rinkinį, būtiną kristalo struktūrai atstatyti naudojant kristalo simetrijos grupės operatorius. Toks aprašymas dažnai nepateikia visos chemikų tikslams reikalingos informacijos; pavyzdžiui, kristale molekulė gali būti aprašyta, nurodant tik jos dalį, o likusi dalis turi būti suskaičiuota, pritaikant informaciją apie simetriją. Toks atvaizdavimo būdas chemikams yra nepatogus ir sukelia sunkumų tolimesnėje duomenų analizėje. Darbe bus siūloma pagaminti iš COD duomenų bazę su pilna informacija apie kristalų elementarius narvelius, t.y. sugeneruojant visus atomus, kurių reikia, norint aprašyti kristalo struktūrą tik elementarių transliacijų pagalba (taip vadinamus P1 narvelius). Šioje duomenų bazėje galima atlikti tolimesnę analizę, pavyzdžiui optinių izomerų paiešką.  EN: In crystallographic structure descriptions, a smallest unique structure fragment – an asymmetric unit – is usually refined and presented; such data set fully describes a crystal structure together with unit cell parameters and crystal symmetry information. The crystallographic description however is not always convenient for chemiinformatic investigations, where researchers are concerned with molecules, their connectivity, proximity and geometry. Thus a full molecular description must be always restored from crystallographic description before further chemoinformatic computations can be carried out. Repeating such computations every time is time and energy consuming, can introduce additional errors and makes a barrier of entry into chemiinformatics research higher for chemists who were not trained in crystallographic computations. In this work we propose to compute a database of full molecule descriptions together with data provenance from the Crystallography Open Database (COD, <https://www.crystallography.net/>), and build a novel open database of chemical structures using models of P1 crystal cells. In this database, investigation of optical isomers (their frequency of occurrence, presence in co-crystal, conformational variability) should be then carried out. | Saulius Gražulis | [grazulis@ibt.lt](mailto:grazulis@ibt.lt) |
| **Srautinio CIF parserio (sintaksinio analizatoriaus) sukūrimas**  **Creating a streaming CIF parser**  Dabar paplitę Crystallography Interchange File (CIF) failų formato sintaksiniai analizatoriai veika, naudodami DOM modelį, t.y. visas failas perskaitomas į atmintį ir tada apdorojamas. Toks metodas visiškai netinka dideliems konkatenuotiems CIF srautams skaityti, pvz. visiems PDB arba COD duomenų bazių įrašams Unix konvejeryje apdoroti. Siūloma sukurti jau esamo cod-tools sintaksinio analizatoriaus pagrindu (C/Bison) srautinį analizatorių, t.y. tokį analizatorių, kuris perskaitytų ir grąžintų failo informaciją po vieną įrašą, ir leistų kreiptis į save daug kartų, pratęsiant sintaksinę analizę nuo tos failo vietos, kurioje buvo sustojęs. Tiriamasis darbo aspektas: darbe bus siekiama palyginti įvairias masinio duomenų apdorojimo architektūras: Unix konvejerius, klasikinius failus, SQL ir NoSQL duomenų bazes, įvertinant jų našumą, universalumą ir patogumą įvairiems mokslininkams kylantiems uždaviniams spręsti. Reikės išnagrinėti teoriškai ir atliekant skaičiavimo eksperimentus, ar srautinį sintaksinį analizatorių galima efektyviai išlygiagretinti daugelio procesorių sistemose, kaip lygiagrečių procesų našumą įtakoja/riboja procesoriaus branduolių skaičius, spartinančios atminties dydis, RAM dydis ir architektūra (lokaliai prieinama atmintis/globali atmintis SMP sistemoje); diskų magistralės pralaidumas. Studentas turės pateikti rekomendacijas, kaip organizuoti kristalografinius skaičiavimus našiausiu būdu ir pademonstruoti tai, naudojant naujai sukurtą srautinį sintaksinį analizatorių.  EN: Current existing CIF parsers use a DOM model, where each CIF *file* is parsed and represented in computer RAM using a DOM object or a parse tree. Such method requires prohibitively much RAM when applied to concatenated streams of CIFs, e.g. for CIF data streams that result from concatenating all records of COD, PDB or other scientific databases. This, application of Unix pipe pattern for processing of crystallographic data is hampered. To make the Unix pipe approach as efficient as it is of line-oriented text processing, a new CIF parser must be created that parses and returned one CIF data block at a time, resuming parsing from the previous location when the client processes the current record and reactivates the parser to deliver the next one. Different methods for construction of such parser (co-routines, callbacks, virtual methods) should be investigated; a parser using the chosen methods should be built. Investigations of parallel execution of the parser and its clients should be performed. The approach should be compared with other data management techniques, such as SQL and noSQL databases, sharded databases, simple file trees, network file systems, version control repositories. Recommendations regarding most efficient computation organisation for scientific computing should be provided. | Saulius Gražulis | [grazulis@ibt.lt](mailto:grazulis@ibt.lt) |
| **Kokybiško CIF parserio (sintaksinio analizatoriaus) sukūrimas Java ir C# (.NET) platformoms**  **High quality CIF parser for Java and .NET platforms**  CIF (Crystallography Interchange Format) yra paplitęs struktūrizuotas duomenų formatas kristalografinių duomenų archyvavimui ir mainams. Yra parašyti kokybiški sintaksiniai šio formato analizatoriai C, Perl ir Python kalbomis. Deja, Java ir .NET aplinkoms šie analizatoriai nėra labai tinkami, nes minėtos platformos sukelia daug keblumų, paleidžiant joms svetimomis kalbomis parašytas bibliotekas. Pvz. Java JNI sąsaja sunkiai perkeliame į mobilias platformas ir dėl to programuotojai labiau vertina „gryna Java“ ("pure Java") parašytus modulius. Darbe bus siūloma perkelti CIF (CIF 1.1 ir CIF 2.0) standarto sintaksinius analizatorius į grynos Java ir C# terpes. Tiriamasis darbo aspektas: bus siekiama įvertinti įvairių sintaksinio analizatoriaus prijungimo prie Java platformos architektūrų bruožus: įtaką programos našumui, poreikius RAM ir CPU laikui; sintaksinio analizatoriaus darbo integravimo į sistemą darbo sąnaudas; programos lydėjimo darbo sąnaudas. Būtų išnagrinėti bent šie bibliotekos prijungimo būdai: a) žemo lygio mašinos kodo bibliotekos prijungimas prie Java aplikacijos JNI metodu ir analogiškais .NET metodais; b) kryžminis Java/.NET bibliotekų kvietimas (naudojant Java VM .NET aplinkoje ir atvirkščiai); c) išorinio proceso iškvietimas d) vien tik Java (angl. „pure Java“) ar vien tik .NET (C#) realizuotos bibliotekos. Bus išnagrinėtas minėtų metodų tinkamumas Web aplikacijoms, mobilioms aplikacijoms Android aplinkoje ir desktop aplikacijoms Linux, Windows, MacOS ir, jei bus galimybė, kitose atvirose OS (FreeBSD, NetBSD, OpenBSD, ReactOS).  Pastaba: čia iš tiesų kalba galėtų eiti apie du panašius darbus ir dvie temas – vieną skirta Java platformos analizei ir pritaikymui, o kitą – .NET platformai; manau, kad minėta analizė ir analizatoriaus įgyvendinimas užims ganėtinai daug laiko, kad pateisintų apsiribojimą viena platforma.  EN: CIF (Crystallographic Interchange Format) is a wide spread standard data exchange format in crystallography, material science, structural biology and adjacent disciplines. CIF uses a structured free format text representation of data, describing file syntax using Backus-Naur forms. Thus a parser is needed to read data form CIF files correctly. Several parsers exist for C, C++, Perl and Python languages; unfortunately, no pure Java or C#/.NET parser is available at the time, which hampers use of CIF in numerous Java and C# chemoinformatics libraries. A native C code could of course be used for Java using JNI interface (and the corresponding binding library for .NET), but such native plug-ins make porting software to different platforms (e.g. mobile computing platforms) more difficult, and developers usually avoid using such libraries, resorting to custom written “pure Java” or “pure .NET” code. The code, unfortunately, is often simplified and not strictly adhering to CIF standard, thus compromising compatibility and data exchange between programs.  In this work, we suggest to write a high-quality pure Java or C# CIF parser, using the IUCr reference grammar, a parser generator, and extensive test sets to check the parser on different CIF syntax cases. The ability of the parser to detect various errors in CIFs should be investigated, as well as its speed and suitability for incorporation into various chemoinformatics libraries (CDK, RDK). | Saulius Gražulis | [grazulis@ibt.lt](mailto:grazulis@ibt.lt) |