# Beszámoló az OTKA F 037567 számú „*Osztott modellek a molekuláris számítástudományban*” című kutatási projekt keretében elért eredményekről

## A kutatás célja, időtartama

A vizsgálódások tárgyai olyan, biokémiai folyamatokat modellező vagy biokémiai folyamatok által inspirált működési elvű számítástudományi eszközök, számítási modellek voltak, melyek fő jellemzője az osztott és párhuzamos működés. A projekt célja volt a molekuláris számítások természetének, a modellek sajátosságainak jobban megfelelő szempontok figyelembe vétele, ezáltal esetleg a biokémiai folyamatok jobb megértése, illetve a formális nyelvek elméletének továbbfejlesztése, eszköztárának bővítése a biokémiai folyamatok és az osztott modellek által inspirált irányba.

A kutatás időtartama 4 év volt, mely a terv szerint 2002. február 1-től 2005. december 31-ig tartott volna, azonban a kutatási szerződés 7.4 pontja alapján, az OTKA bizottság elnökének engedélyével a munkát 2005 július 1-től egy évre felfüggesztettem, így a projekt záró dátuma 2006. december 31-re módosult. A felfüggesztésre irányuló kérelmem alapja az volt, hogy 2005 tavaszán elnyertem a német Alexander von Humboldt alapítvány kutatási ösztöndíját és így a 2005-ös év második valamint a 2006-os év első félévét Magyarországon kívül, a magdeburgi Otto-von-Guericke Egyetem elméleti számítástudományi csoportjának vendégkutatójaként töltöttem el.

## A DNS rekombináció motiválta számítási eszközök

A biokémiai folyamatok tanulmányozása nem idegen az elméleti számítástudomány, illetve az automaták és formális nyelvek elméletének tudományágától, hiszen a DNS molekulák négy alkotóelemből felépülő láncai természetes módon feleltethetők meg egy négy szimbólumot tartalmazó ábécé betűiből képzett jelsorozatoknak, szavaknak, a molekulák halmazai pedig szóhalmazoknak, nyelveknek.

Az utóbbi években a kémiai reakciók modellezésére a molekulák bomlását és összekapcsolódását leíró új nyelvi műveletek születtek, illetve rájuk alapozva osztott kiszámítási eszközök is megjelentek. Az ilyen eszközök egyik lehetséges alapgondolata a következő. Legyen adva egy szósokaság valamint a szavakon értelmezett műveleteknek valamilyen halmaza és tekintsük az eszköz által kiszámított (generált) nyelvnek azon szavak összességét, amelyek a kiindulási sokaságból a műveletek (akár iterált) alkalmazása segítségével megkaphatók.

A sodráson (splicing), vagyis a DNS szálak rekombináns viselkedésének utánzásán alapuló nyelvleíró, illetve kiszámítási eszközök tanulmányozására az elmúlt években

intenzív kutatások irányultak. A témakörrel kapcsolatos eredményekről összefoglaló olvasható a [2.] könyvfejezetben és az [5.] monográfiában. A sodrási rendszerek eddig vizsgált változatai azonban két fontos szempontot általában figyelmen kívül hagynak.

Először is, a molekulák sokaságát szavak (matematikai értelemben vett és nem feltétlenül véges) halmazaival modellezik, amik tehát nem veszik figyelembe az egyes szimbólumsorozatok előfordulásának gyakoriságát, hiszen minden szóból potenciálisan végtelen számú másolat áll rendelkezésre, vagy ha multihalmazokkal dolgoznak is, azaz az egyes szavakból rendelkezésre álló példányok száma korlátozott lehet, mégis alapvetően fontos követelmény az, hogy bizonyos szavak számossága végtelen lehessen. (Lásd a [2.] könyvfejezet 3.3 alfejezetét, illetve az [5.] monográfia 8.6 fejezetét.)

A második szempont a műveletek párhuzamos végrehajtását érinti. A modellezni kívánt molekulasokaság alapvető tulajdonsága, hogy a biokémiai reakció párhuzamosan zajlik, a fent vázolt modellekben azonban, mivel a szavakból tetszőleges számú másolat áll rendelkezésre, a szósokaság evolúciója gyakorlatilag „szekvenciális" folyamat illetve akár ilyennek tekinthető, hiszen a két működési mód (szekvenciális és párhuzamos) eredményében különbség nem mutatható ki.

A fentiek miatt kezdtük el Jürgen Dassowval, a magdeburgi (Németország) egyetem professzorával közösen a véges multihalmazokkal (véges számú elem véges számú példányban) dolgozó rendszerek vizsgálatát ahol a műveletek párhuzamosan zajlanak és a szavakat „elhasználják," azaz egy-egy művelet végrehajtása során a szavak amelyekre a műveletet alkalmaztuk eltűnnek, míg az eredményként kapott egy vagy több szó megjelenik a molekulasokaságot reprezentáló multihalmazban. Mivel egy szó egyszerre csak egyetlen műveletben vehet részt és a „semmiből" új szavak nem keletkezhetnek, az ily módon definiált rendszerek véges nyelveket generálnak (ezért a számítási erő helyett más jellegű kérdések vizsgálata érdekes), valamint módot adnak a műveletek párhuzamos végrehajtásának többféle módon való definíciójára is.

A jelentéshez tartozó közleménylistán 4.-ként szereplő dolgozatban megvizsgáltuk a szekvenciálisan, valamint két különböző módon definiált párhuzamosság alapján működő sodrási rendszereket (multiset splicing systems), és összehasonlítottuk az ily módon kiszámítható nyelveket. A nyelvek meghatározására is több módszert vezettünk be, vizsgáltuk a multihalmaz nyelvet, azaz egy kiindulási állapotból létrehozható különböző multihalmazok összességét, valamint a szavak nyelvét, azaz az előálló multihalmazokban előforduló, vagyis a valamilyen módon előállítható szavak halmazát. A szavak nyelveivel kapcsolatban megmutattuk, hogy gyakorlatilag minden $n$ természetes számhoz létezik olyan nyelv, ami $n$-nél kisebb számosságú kiindulási multihalmazzal rendelkező rendszer segítségével nem állítható elő, a multihalmaz nyelvekről pedig bebizonyítottuk, hogy az $n$ számosságú kiindulási multihalmazból háromféle módon (szekvenciálisan, illetve kétféle párhuzamos alkalmazás segítségével) előállítható nyelvek egymással nem összehasonlítható nyelvosztályokat képeznek.

A közleménylista 10. dolgozatában is hasonló elvek alapján konstruált rendszereket vizsgáltunk, ezek azonban a sodrás helyett a csillós egysejtűek (ciliates) genetikai

anyagának rekombinációját modellező műveleteket használnak. A csillós egysejtűek olyan egyszerű egysejtű organizmusok, melyekben a DNS molekulák rekombinációjának folyamata viszonylag részletesen ismert, tehát a különféle matematikai modellek pontossága jól ellenőrizhető. Az utóbbi években számos, a csillós egysejtűek örökítőanyagának rekombináns viselkedését leíró formális nyelvi művelet született (lásd pl. az [1.] monográfiát), vizsgálódásaink során mi is ezekből indultunk ki.

Itt is sodrási rendszerekben bizonyított összehasonlíthatatlansági tételekhez hasonló eredményeket kaptunk, de nem csak a multihalmaz, hanem bizonyos esetekben a szavak nyelveire is. Ezeknél az eredményeknél azonban még érdekesebb lehet a rendszer azon tulajdonsága, hogy bár a rekombinációs műveletek megfordíthatók (reversible), azaz egy művelet alkalmazásának eredményeként kapott szavakból mindig létrejöhetnek a kiindulási szavak is (a művelet megfordítása is mindig alkalmazható), a párhuzamosság és a rendelkezésre álló szavak véges száma miatt azonban a rendszer által felvett állapotok sorozata mégsem „megfordítható," azaz egy-egy a rendszer fejlődése során elért konfigurációból nem feltétlen vezet visszaút a korábbi konfigurációk felé.

Ennek a nem megfordítható fejlődésnek a megjelenése azért fontos, mert választ jelenthet arra a kérdésre, hogy a külön-külön megfordíthatónak tetsző rekombinációs műveletek segítségével létrejövő, a csillós egysejtűek genetikai anyagának felépítését eredményező folyamat hogyan lehet mégis egyirányú. Ez a kérdés másokat is foglalkoztat, egy lehetséges megoldás található pl. a [8.] dolgozatban. (A szerzők javaslata szerint a rekombináció folyamatában nem csak a rekombinálandó szó vagy szavak vesznek részt, hanem egyfajta katalizátorként az esetleg más forrásból már létező eredmény is, azaz a rekombináció során olyan szavak jönnek létre nagyobb valószínűséggel amilyenek már eleve léteznek.)

## Membrán rendszerek

A membrán rendszerek (P rendszerek) az élő sejtek működésének modellezésén alapuló kiszámítási eszközök. Fő részei a hierarchikusan elrendezett, egymást tartalmazó membránok, amelyek különböző tulajdonságokkal bíró régiókat zárnak magukba. A régiók leírására az általuk tartalmazott objektumok multihalmaza és az objektumok egymás közti kölcsönhatását leíró szabályok szolgálnak. A régiók közti kommunikáció a membránokon keresztül valósul meg, ezeket a membránok tulajdonságait megfogalmazó szabályok írják le. 1998 óta a membrán rendszerek kutatása a molekuláris, nem-konvencionális számítástudomány sikeres ágává vált. (Az amerikai Institute for Scientific Information (ISI) a számítástudomány területén 2003 februárjában Gh. Paun dolgozatát, [6.], mely a terület fejlődését elindította „fast breaking paper"-ként, magát a tudományterületet 2003 októberében „emerging research front"-ként értékelte. A membrándszerekről lásd a [7.] monográfiát.)

A P rendszerek területén folytatott kutatásaim több témakört érintettek. Az ún. párhuzamosan újraíró P rendszereket (parallel rewriting P systems) Daniela Besozzival (Universita degli Studi di Milano, Milánó, Olaszország), Giancarlo Maurival, és Claudio

Zandronnal (Universita degli Studi di Milano-Bicocca, Milánó, Olaszország) vizsgáltuk. A párhuzamosan újraíró P rendszerekben a membránokban lévő objektumokat szimbólumsorozatok, azaz szavak reprezentálják, az objektumok változásait pedig olyan produkciós szabályok írják le, melyek alkalmazása párhuzamosan, azaz egy lépésben a szavak összes szimbólumára vonatkozóan egyszerre történik. Mivel azonban az egyes produkciós szabályok nem csak a szimbólumok átírását, hanem az átírás után keletkező szót egy membránon keresztül valamelyik szomszédos régióba át is küldhetik, előfordul, hogy az egy lépésben párhuzamosan alkalmazandó szabályok konfliktusban vannak egymással (target conflict), melyeknek a feloldására társszerzőim különböző módszereket vezettek be. A közleménylista 3. dolgozatában leírt közös kutatásaink során azt vizsgáltuk, hogy az egyes konfliktusfeloldó módszerek alkalmazásával kiszámítható nyelvosztályok milyen viszonyban vannak egymással, illetve a Lindenmayer rendszerek (klasszikus párhuzamos szabályalkalmazással működő újraíró rendszerek) különféle változatainak segítségével leírható nyelvosztályokkal

A membrán rendszerek egy másik változata, a symport/antiport rendszerek olyan membrán rendszerek, amelyekben csak az objektumok mozgása megengedett, azok evolúciója, változása nem. Az objektumok mozgását az egyes membránokhoz rendelt *(x,be; y,ki)* alakú szabályok írják le, ahol *x* és *y* objektumok mutihalmazait jelöli. A fenti szabály szerint *x* multihalmaz objektumai behatolhatnak a membrán által körbezárt régióba, miközben (ugyanebben a lépésben) *y* multihalmaz objektumai elhagyják azt. Ha az objektumok bizonyos további feltételeknek megfelelő módon, de alapvetően a fent vázolt típusú szabályok párhuzamos alkalmazása alapján vándorolnak a régiók között, illetve a rendszer környezetéből tetszőleges mennyiségben behatolhatnak a külső membránon keresztül, akkor a symport/antiport rendszerek képesek természetes számok minden rekurzíve felsorolható halmazának előállítására. (A kiszámított értéket a működés végén egy bizonyos régióban található objektumok száma adja meg.)

Érdekes kérdés ezzel kapcsolatban, hogy mennyire bonyolult kommunikációs szabályokra van szükség, illetve milyen kapcsolat van a szabályok és a membránstruktúra szükséges bonyolultsága között. A problémával többen foglalkoztak, míg 2004-ben sikerült megmutatni, hogy a maximális kiszámítási erő a lehető legegyszerűbb kommunikációs szabályokkal (amikor *x* és *y* is csak egyetlen objektumot tartalmaz) és kilenc membránnal is elérhető. A kérdéskör vizsgálata tovább folyt, még ugyanebben az évben különböző szerzők egymást sorban követő eredményei szerint a szükséges membránok számát hatra, ötre, majd négyre sikerült csökkenteni. Nekem a közleménylista 6. dolgozatában ezt a számot, az akkor ismert legjobb korlátot sikerült tovább, háromra csökkentenem.

Hasonló kérdéseket vizsgáltunk Csuhaj Varjú Erzsébettel közösen az ún. sarjadzó (gemmating) rendszerek bizonyos változataival kapcsolatban is, melyekben az objektumokat sztringek reprezentálják, az objektumok változását környezetfüggetlen újraíró szabályok, régiók közti mozgását pedig ún. pre-dinamikus szabályok írják le. A pre-dinamikus szabályok olyan speciális környezetfüggetlen szabályok, melyek alkalmazása után az átírt szó egy adott indexű másik régióba kerül. Egy 2003-as eredmény szerint nyolc membránnal minden rekurzíve felsorolható nyelv előállítható

ilyen rendszerekkel, ha a számítás eredményének a külső membránból a környezetbe kilépő szavak halmazát tekintjük, mindehhez kilenc membrán szükséges akkor, ha a rendszer csak pre-dinamikus szabályokat használ. A közleménylista 5. dolgozatában a szükséges membránok számát javítottuk háromra illetve négyre, valamint azt is megmutattuk, hogy a három membrános korlát optimális.

A fentiek mellett, melyek a P rendszerekkel kapcsolatos kutatások hagyományosnak tekinthető irányaiba esnek, Csuhaj Varjú Erzsébettel és Gheorghe Paunnal (University of Sevilla, Spanyolország és a Román Tudományos Akadémia Matematikai Intézete) közösen a membrán rendszerek és a grammatika-rendszerek bizonyos elemeinek összekapcsolásával létrejövő új kiszámítási modelleket is vizsgáltunk. (A grammatika-rendszerek tárgykör a formális nyelvek elméletének elismert területe, melynek elindítása Csuhaj Varjú Erzsébet és Jürgen Dassow nevéhez köthető.) Az egyik ilyen modell a szövetszerű (tissue) P rendszerek egy módosított változata, melyben a membránokban lévő objektumokat sztringek reprezentálják és a régiók közötti kommunikáció ún. igény szerinti módon történik. A szövetszerű P rendszerek fontos jellemzője, hogy a membrán struktúrát leíró gráf nem feltétlenül fa, azaz a régiók közötti kapcsolat nem csak a tartalmazás (tartalmazottság) reláció lehet. Ez a tulajdonságuk feltűnő hasonlóságot mutat az ún. párhuzamos, kommunikáló (PC) grammatika-rendszerek komponensei közötti kapcsolatokkal ami a PC grammatika-rendszerek kutatásai során kidolgozott módszereket alkalmazhatóvá teszi a szövetszerű P rendszerek említett variánsainak vizsgálata során is. A közleménylista 12. közleményében szereplő eredményeink azt mutatják, hogy ezek a rendszerek képesek tetszőleges rekurzíve felsorolható nyelv generálására, illetve, mivel a régiókban lévő szavak száma exponenciális sebességgel is képes növekedni, alkalmasak bizonyos NP teljes problémák polinomiális időben való megoldására is.

Jozef Kelemen és Alica Kelemenová (University of Opava, Csehország) további társszerzőkkel együtt tanulmányoztuk a korábban bevezetett (lásd [4.]) P kolóniák tulajdonságait is. Hasonlóan a grammatikarendszerek elméletéből ismert kolóniához, a P kolónia is egészen egyszerű, egymással csak a közös környezet megváltoztatása útján kommunikáló ágensek közössége, melyek leírására ezúttal nagyon egyszerű membránrendszereket használunk. A közleménylista 8. dolgozatában közölt eredményeink azt mutatják, hogy az ily módon szervezett ágens-közösség számítási ereje meglepően nagy, azaz a közösség viselkedése emergensnek tekinthető.

## P automata

A Csuhaj Varjú Erzsébettel közösen bevezetett P automata olyan membrán rendszer, amelyben az objektumok nem lépnek reakcióba egymással, azaz a kiszámítás folyamata csak az egyes régiók közötti kommunikáció segítségével valósul meg, a számítás eredményét pedig a rendszer működése során a külső membránon keresztül a rendszerbe behatoló objektumok sorozatának segítségével határozzuk meg, hasonlóan ahhoz, ahogy a hagyományos értelemben vett automaták által kiszámított nyelvet is a bemeneten lévő szavak, jelsorozatok végigolvasása határozza meg. A fogalom a kutatói közösségben élénk visszhangot keltett, amit az is jelez, hogy a közleménylistában 2. helyen szereplő

dolgozatunkra, (és annak a konferenciakiadványban közölt első változatára amit a lista nem tartalmaz) rövid idő alatt 41 hivatkozás történt.

A közleménylista 11. dolgozatában azt a problémát vizsgáltam, hogy lehetséges-e pusztán szintaktikai megszorításokkal (tehát pusztán a rendszer működését vezérlő szabályok alakjára kimondott korlátozások segítségével) valamilyen már ismert nyelvosztályt, esetünkben a környezetfüggetlen nyelvek osztályát, karakterizálni.

A P automaták egyik rendkívül érdekes tulajdonsága, hogy mivel a számítás eredményét a rendszer működése során a külső membránon keresztül behatoló objektum-multihalmazok sorozatának segítségével határozzuk meg, és mivel az egyes kiszámítási lépésekben az automata által beolvasható egymástól különböző objektum-multihalmazok száma nem feltétlenül véges. Ebből következően természetesen adódik a lehetőség, hogy a P automatát végtelen ábécé felett értelmezett nyelvek leírására használjuk fel. A végtelen ábécé feletti nyelvek vizsgálatának igénye olyan gyakorlati kérdésekkel kapcsolatban merül fel, mint az adatbázis lekérdezések, az interneten történő navigálás, vagy az XML dokumentumjelölő (markup) nyelv, vagyis amikor a vizsgálandó nyelv szavaiban mint jelsorozatokban szereplő jelek értékkészlete nem egy előre adott véges halmaz. A „hagyományos” megközelítés során (lásd pl. [3.]), a végtelen ábécé feletti nyelveket leíró automaták utasításkészletét (vagy a grammatikák újraíró szabályait) olyan műveletekkel egészítik ki, amelyek lehetőséget adnak végtelen számú szimbólum kezelésének véges számú utasítással történő leírására. Ez általában olyan képességekkel ruházza fel a kérdéses eszközöket, amelyek azokat nehezebben kezelhetővé teszik. Az közleménylista 9. dolgozatában Jürgen Dassowval közösen bevezetett P véges automata ezzel szemben egy olyan „meta” szabályrendszer illetve utasításkészlet, amely véges számú különböző objektummal dolgozik, azaz véges számú utasítással specifikálható, és mégis, a potenciálisan korlátlan számú bemenet segítségével bizonyos szempontból „természetes” általánosítását nyújtja a véges automata illetve a reguláris nyelv fogalmának.

A P véges automata olyan P automata, melynek szabályai bizonyos formai megszorításoknak eleget téve nagyon leegyszerűsített működést produkálhatnak csupán. Ez a leegyszerűsített működés azonban elegendő ahhoz, hogy amint azt megmutattuk, a véges ábécé feletti nyelvek közül a P véges automata pontosan a „hagyományos” értelemben reguláris nyelveket írja le, míg végtelen ábécék felett egy olyan nyelvosztályt karakterizáljon, amely nem túlságosan bonyolult, és mégis magába foglal két, más szerzők által korábban a reguláris nyelvek általánosításaként javasolt (lásd [3.] és [9.]), egymással egyébként össze nem hasonlítható végtelen ábécé feletti nyelvosztályt.

## Hivatkozási jegyzék

[1.] Ehrenfeucht, T. Harju, I. Petre, D. M. Prescott, G. Rozenberg. *Computation in Living Cells. Gene Assembly in Ciliates*. Springer Verlag, Berlin Heidelberg, 2004.

[2.] T. Head, Gh. Paun, D. Pixton. Language theory and molecular genetics: Generative mechanisms suggested by DNA recombination. In: G. Rozenberg, A. Salomaa (editors) *Handbook of Formal Languages*, volume II, chapter 7, Springer Verlag, Berlin Heidelberg, 1997, pages 295-360.

[3.] M. Kaminski, N. Francez. Finite-memory Automata. *Theoretical Computer Science* 134 (1994) pages 329-363.

[4.] J. Kelemen, A, Kelemenová, Gh. Paun. Preview of P colonies. A bio-chmically inspired computing model. In: M. Bedau et al. (editors) *Ninth International Conference on the Simulation and Synthesis of Living Systems, Alife IX., Workshop and Tutorial Proceedings*. Boston MA, pages 82-86.

[5.] Gh. Paun, G. Rozenberg, A. Salomaa. DNA *Computing. New Computing Paradigms*. Springer Verlag, Berlin Heidelberg, 1998.

[6.] Gh. Paun. Computing with membranes. *Journal of Computer and System Sciences*, 61 (2000), pages 108-143.

[7.] Gh. Paun. *Membrane Computing. An Introduction*. Springer-Verlag, Berlin Heidelberg, 2002.

[8.] D. M. Prescott, A. Ehrenfeucht, G. Rozenberg. Template-guided recombination for IES elimination and unscrambling of genes in stichotrichous ciliates. *Journal of Theoretical Biology* 222 (2003) pages 323–330.

[9.] Otto, F.: Classes of Regular and Context-free Languages over Countably Infinite Alphabets. *Discrete Applied Mathematics* 12 (1985) pages 41-56.