Magyar Tudomány 179(2018)5, 636–642
DOI: 10.1556/2065.179.2018.5.6

SZAGÉRZÉKELÉS PREDIKCIÓJA GÉPI TANULÁS ALKALMAZÁSÁVAL

PREDICTION OF OLFACTORY PERCEPTION WITH MACHINE LEARNING

Turu Gábor[1], Cserző Miklós[2], Szalai Bence[3], Hunyady László[4]

[1]PhD, adjunktus, turu.gabor@med.semmelweis-univ.hu
[2]PhD, tudományos főmunkatárs
[3]PhD, tanársegéd
[4]az MTA rendes tagja, PhD, egyetemi tanár
Semmelweis Egyetem Élettani Intézet

ÖSSZEFOGLALÁS

Míg a legtöbb szenzoros működés esetén az inger és a kialakult érzet tulajdonságai közötti összefüggések már régóta ismertek (például: fény-hullámhossz – szín, hangfrekvencia – hangmagasság), a szaglás esetén nem állt rendelkezésünkre olyan általános modell, amely segítségével egy szaganyag struktúrája alapján megmondható lenne annak illata.

A DREAM Olfaction Predicting Challenge során a Rockefeller University Smell Study még nem publikált kísérletsorozatának eredményeit felhasználva kellett olyan modellt kidolgozni, mely a molekuláris struktúra alapján prediktálja szaganyagok illatát.

Munkacsoportunk a Dragon-deszkriptorokat és a Morgan-fingerprinteket felhasználva dolgozott ki egy többszörös lineáris regresszió alapú modellt.

Az általunk kidolgozott modell a szaganyagok intenzitását 0,71/0,53 (populációs/egyéni predikció), kellemességét 0,58/0,34, a további minőségi tulajdonságokat átlagosan 0,53/0,21 korrelációval prediktálta, ezzel az egyéni predikcióban 2/18, a populációszintű predikcióban 7/19 helyezést érve el. Tekintettel arra, hogy a szaglás hátterében egy ligand-receptor kölcsönhatás áll, elképzelésünk szerint módszerünk felhasználható lehet (nem szagló) receptorok ligandkötésének prediktálására is.

ABSTRACT

While for most of the sensory functions relations between signal and perception is well known (i.e. wavelength and colour of light, frequency and tone of the sound) in case of odour perception we have no general model for prediction of a smell based on the chemistry of a given compound.

The DREAM Olfaction Predicting Challenge aimed to develop such a "smell from chemical formula" model based on the Smell Study dataset of the Rockefeller University.

Our group developed and tested a multiple linear regression model using the combination of Dragon descriptors and Morgan features of molecules in a random forest machine learning algorithm.

Our model reached 0.71/0.53 prediction score for smell intensities (population level/per person level), 0.58/0.34 prediction score for pleasantness (population/per person), the average

score for all the tested properties is 0.53/0.21 for population level and per person level respectively. In the final ranking our model scored as 2/18 for per person test and 7/19 for the population level test. Considering that the odour perception is based on a ligand – receptor interaction the applied methods can be generalized for non-smell related cases of receptor involved signalling.

Kulcsszavak: szaglás, receptor-ligand kölcsönhatás, gépi tanulás, in silico screening, DREAM Challenge

Keywords: smell, receptor-ligand interaction, machine learning, in silico screening, DREAM challenge

Az elmúlt évtizedekben a biológiai tudományokban történt fejlődés egyik következménye a kísérleti adatok mennyiségének exponenciális méretű növekedése. Az adatok mennyiségének növekedése szükségessé teszi azok számítógépes analízisét, hogy jobban megérthessük az adatok mögött rejlő összefüggéseket, biológiai működéseket. A számítógépes analízis egyik módja a gépi tanulás. Ennek során a számítógépes programnak ismert példákat mutatunk, amelyek alapján önállóan tanulja meg a szabályokat, és képes nem ismert mintákat osztályozni vagy hozzájuk értékeket társítani. A legegyszerűbb gépi tanulásnak tekinthetjük az általánosan ismert lineáris regressziót, ahol is egy adatsor néhány x, y pontja alapján következtethetünk további x-ekhez tartozó y értékekre. Az elmúlt évtizedekben jelentős fejlődésen ment keresztül a tudományág, és ma már számos területen alkalmazzák a módszert, az e-mailek spamszűrésétől kezdve az önjáró autók környezetfelismeréséig. A biológiában több fontos alkalmazási területen van jelentősége: adatok automatizált elemzése, nem mért adatok prediktálása és a mért adatok mögötti összefüggések, mechanizmusok feltárása. Az automatizált adatelemzés segíthet olyan feladatok gyors és elfogulatlan elemzésében, amelyek egyébként jelentős szakképzett humán erőforrás munkaidőt igényelnek, mint például szövettani metszetek elemzése vagy röntgenfelvételek diagnosztizálása. Másrészt, amennyiben egy biológiai rendszerről számos paraméter áll rendelkezésre (tipikusan például DNS- és RNS-szekvenálási eredmények), a gépi tanulás módszerével kiválaszthatók lehetnek azok, amelyek ténylegesen számítanak a vizsgált jelenség, például tumor kialakulása hátterében, így betekintést nyerhetünk egyes jelenségek mögött megbúvó összefüggésekbe.

A mért biológiai adatok helyes értelmezését nehezítheti, hogy az összes mért adat ismeretében az eredményeket jól leíró modell nem feltétlenül általánosítható, alkalmazható új mérési adatokra. Másrészt, nem feltétlenül a méréseket végző szakemberek rendelkeznek azzal a szaktudással, amelyek az adatok értékeléséhez szükségesek. Egy lehetséges megoldás ezekre a problémákra a *crowdsourcing,* vagyis az adatok elemzésének kiszervezése független személyek számára, akik

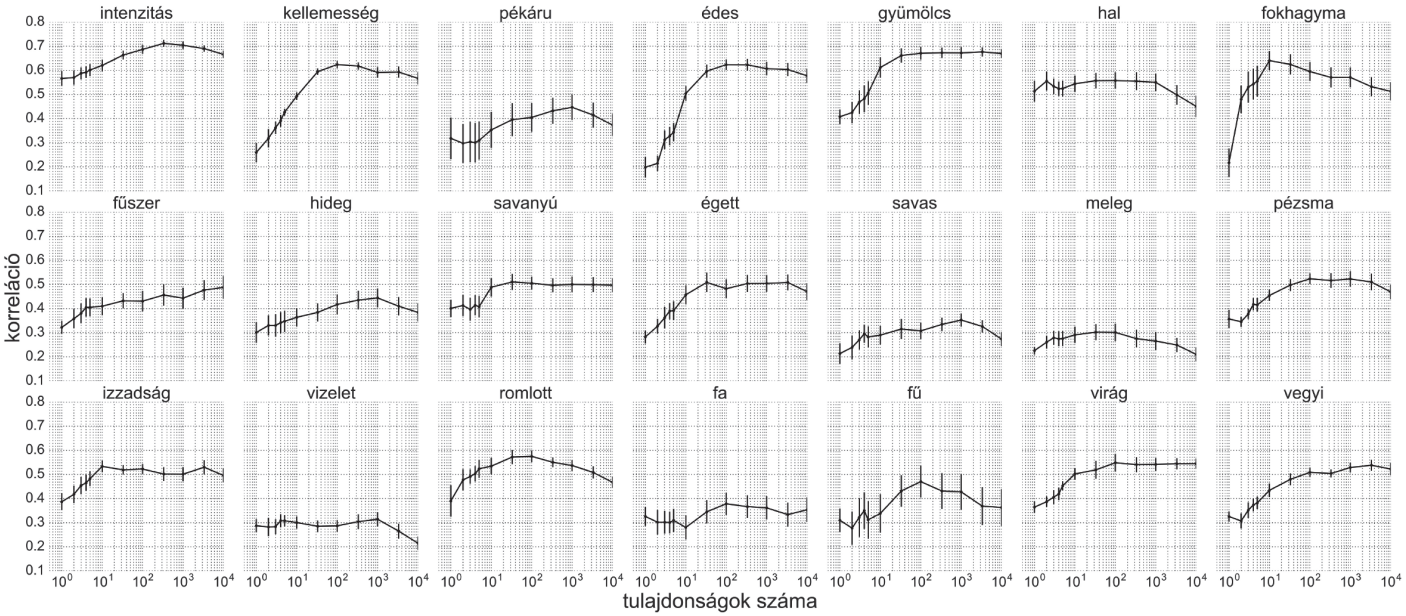
nem elfogultak az adatokkal szemben, és nem rendelkeznek olyan elvárásokkal, amelyek akaratlanul is befolyásolhatják az adatok elemzését és a végső konklúziót. Ennek során a méréseket végző szervezet jellemzően *online* felületen megosztja az adatait, amelyekkel majd a terület iránt érdeklődő adatelemzők dolgoznak.

Egy ilyen megoldást kínál a DREAM Challenges (Dialogue for Reverse Engineering Assessments and Methods) nonprofit szervezet, amely orvosi és biológiai kérdések megoldására szervez versenyeket. A versenyekhez kutatást végző csoportok vagy szervezetek biztosítják az adatokat, és az elemzés kiszervezését online adatelemző versenyek formájában a DREAM biztosítja. A versenyek hasznosak az adatokat szolgáltató csoportnak, mert sokan (10–100 csapat) dolgoznak az adatain, és új ötletek születhetnek, valamint jelentősen nagyobb hatékonyság, gyorsabb eredmény várható. Másrészt jó a verseny az adatelemzőknek, mert valós, érdekes, igazi adatokhoz férnek hozzá, kollaborációk születhetnek különböző területekről érkezőkkel és a versenyek legjobbjai társszerzői lehetnek nagy presztízsű lapokban megjelent publikációknak. Ez a megoldás jó a tudományterületnek is, hiszen az adott területtől távolabb álló szakembereket is bevon a kutatásba, ami a különböző szemléletekből adódóan új ötletek születését segíti. Az átlátható, mindenki számára megismerhető módszertanból született modellek viszonyítási alapot jelentenek a későbbi kutatások irányának meghatározása szempontjából.

Munkacsoportunk a DREAM Olfaction Prediction Challenge-ben vett részt. A verseny során arra a kérdésre kerestük a választ, hogy mely molekuláris struktúrák határozzák meg a különböző szagok intenzitását, kellemességét és a szag minőségét. A verseny célja olyan modellek megépítése, amelyek a molekulák szerkezetéből a lehető legpontosabban megjósolja azok szagát. A látás és hallás esetében már korábban pontosan ismert volt, hogy az ingerek (hang, fény) milyen fizikai tulajdonságai vezetnek a különböző szín és hangmagasság érzékeléséhez. Szag esetében is ismert, hogy különböző molekulák más és más szagérzetet váltanak ki, de hogy pontosan mely molekuláris struktúrák milyen érzethez vezetnek, azt eddig nem térképezték fel. A Rockefeller University kutatócsoportja egy 49 (nem szakértő) személlyel tesztelt 476 különböző molekulát két különböző koncentrációban. A tesztalanyok feladata az volt, hogy 100 pontos skálán pontozzák a szagok intenzitását, kellemességét, és azt, hogy mennyire sorolható be a szag 19 különböző illattípusba (pékáru, édes, gyümölcs, hal, fokhagyma, fűszer, hideg, savanyú, égett, savas, meleg, pézsma, izzadság, vizelet, romlott, fa, fű, virág, vegyi) (Keller–Vosshall, 2016). A munka során a kutatócsoport által gyűjtött adatok alapján a versenyzők előbb külön-külön, majd a legjobban teljesítők közösen, kollaborációban egy olyan számítógépes modellt állítottak fel, amely további nem tesztelt molekulák szagát képes megjósolni. A predikcióhoz a molekula képlete, szerkezete állt rendelkezésre, illetve a szerkezetből számítógépes program által számított jellemzők (a használt szoftver alapján elnevezett DRAGON-tulajdonságok). A verseny során szabadon lehetett a molekulából származtatott egyéb

jellemzőket is használni. A 476 mintából első körben 338-at, majd további 69-et osztottak meg a versenyzőkkel (tréningszett). A visszatartott 69 molekula szagát kellett megjósolni a verseny végső fázisában az első, 338 molekula tulajdonságai és kísérleti adatai alapján készült modell segítségével. A modelleket két kategóriában kellett elkészíteni: egyikben egyének szintjén, másikban populációs szinten kellett a molekulák szagát jósolni. Az értékelés alapja a prediktált és a valós adatok Pearson-korrelációja volt.

Saját predikciónkhoz a DRAGON-jellemzőkön kívül úgynevezett Morgan-ujjlenyomatokat, pontosabban a Morgan-ujjlenyomatok alapján számolt molekulahasonlóságot is felhasználtuk (Rogers–Hahn 2010). A Morgan-ujjlenyomatok meghatározása során molekularészletek meglétét keressük a vegyületekben, ami alapján egyedi mintázatot kapunk minden egyes vegyületre. Ezen mintázatok alapján meghatározhatjuk, hogy két molekula milyen mértékben hasonlít egymásra. Feltételezésünk szerint strukturálisan hasonló molekulák hasonló szagérzetet képesek létrehozni. Minden szaganyag esetén tehát meghatároztuk, mennyire hasonlítanak a többi 476 molekulára, illetve mennyire hasonlítanak további 1961 ismert illatú vegyületre. A hasonlósági mintázat a hipotézisünk szerint jellemző lehet egy-egy adott szagra. A hasonlósági mintázat és a DRAGON-jellemzők adták azokat a tulajdonságokat, amelyek alapján a modellünket illesztettük. Predikciós modellként lineáris regressziót használtunk annak egyszerűsége és gyorsasága miatt. A kódoláshoz Python programozási nyelvet és sklearn- (Pedregosa et al., 2011), valamint rdkit- (URL1) könyvtárakat alkalmaztunk. Mivel lineáris regresszió esetén, amennyiben a tulajdonságok száma (jelen esetben több ezer) jelentősen meghaladja a minták számát (ami most 407), fennállt a veszélye annak, hogy a modell túlságosan illeszkedik a tréningadatokra, de nem képes pontosan megjósolni újabb szaganyagokat (*overfitting*nek nevezett jelenség). A jelenség hátterében az áll, hogy minél több tulajdonság alapján illesztjük a modellünket, annál nagyobb az esélye, hogy az algoritmus olyan tulajdonságoknak tulajdonít nagy jelentőséget, amelyek véletlenül csak az adott szagú vegyületeknél vannak jelen, de igazából nincs jelentőségük az adott szagérzet kiváltásában. Hogy ezt elkerüljük, a modell illesztése előtt RandomizedLasso-algoritmussal kiválasztottuk a több ezer jellemző közül azokat, amelyek a legjobb predikciós értékűnek bizonyultak. A kiválasztás során több száz modellillesztést végez az algoritmus úgy, hogy minden esetben véletlenszerűen változik, melyik mintákat veszi be az illesztésbe. A felhasznált jellemzőket aztán a felhasználás gyakorisága alapján sorba tesszük, és a leggyakoribbak alapján végezzük a modell végső illesztését. A különböző szagjellemzők jóslásában nem ugyanolyan számú tulajdonságra volt szükség: fokhagymaillat esetén néhány molekulajellemző már elég a szag meglétéhez, azonban például a szag intenzitásának becslése több száz tulajdonság bevonása után is javítható volt még újabbak hozzáadásával *(1. ábra).*



1. ábra. A predikciók függése a felhasznált tulajdonságok száma alapján. A predikciót különböző számú tulajdonság felhasználásával elvégezve megvizsgáltuk a korrelációt a valós adatokkal. Az értékek átlagos értékek + – szórás tíz keresztvalidációból.

Nem minden szagtípust lehetett nagy pontossággal megjósolni. Véleményünk szerint ebben jelentős szerepe lehet annak, hogy bizonyos szagok nagyon jól meghatározhatók és könnyen értelmezhetők, mint például a fokhagymaszag, míg mások nehezebben azonosíthatók, mint például a meleg szag. Az általunk kapott becslések alapján az egyének szintjén második (18-ból), a populáció szintjén 7. (19-ből) helyezést értünk el. A helyezések alapján meghívást kaptunk a verseny második, kollaborációs szakaszában való részvételre, amikor is a legjobb csapatok közösen készítenek egy predikciós modellt, felhasználva a külön-külön szerzett tapasztalatokat. A kollaborációban végül két modellt készítettünk, egy lineáris modellt, amely megfelelt az általunk használt lineáris regressziónak, és egy nem lineáris, döntési fa alapú modellt (Keller et al., 2017). Mindkét modellben a DRAGON-jellemzőket és a Morgan-hasonlóságokat használtuk fel a tréning során, ugyanis a számos egyéb jellemző közül csak ezek bizonyultak a gépi tanulás számára hasznos információnak. A két modell végül nem tért el jelentősen. Az intenzitást egyéni szinten 0,56, populációra 0,78, a kellemességet 0,41, illetve 0,71, a maradék 19 szagtulajdonságot átlagosan 0,21 és 0,55 korrelációs értékekkel sikerült megbecsülni. Felmerül a kérdés, hogy a kapott korrelációk mit jelentenek, mennyire tekinthetők jónak. Ennek eldöntésére használhatjuk azokat a mintákat, amelyeket a vizsgálatot végző személyek két különböző alkalommal is jellemeztek. A második alkalommal ugyanazokat a vegyületeket kicsit más pontszámokkal értékelték, így ezen tesztelés-újratesztelés (test-retest) vizsgálatokból is számíthatunk korrelációkat. A legjobb predikciók esetén sem számíthatunk arra, hogy a jóslással jobb korrelációkat érünk el, mint ugyanazon személyek ugyanazon minták újraszagoltatásával. Összehasonlítás után az derült ki, hogy a tesztelés-újratesztelés és a predikciók pontossága populációs szinten a legtöbb szagtípus esetén statisztikailag nem különbözött jelentős mértékben, vagyis az alkalmazott modellek hatékonysága közel van az elméletileg maximálisan elérhetőhöz (Keller et al., 2017).

A kapott eredmények jelentősége egyrészt, hogy a szagok molekuláris struktúrájának ismerete hozzájárul a szaglás élettanának jobb megértéséhez, másrészt ipari hasznosításuk is lehetséges új illatok, szagmolekulák megtalálásában, összeállításában vagy még nem tesztelt vegyületek szagának prediktálásában. Véleményünk szerint a módszer használható lehet a gyógyszeriparban is új hatóanyagok megtalálásában. A szagok úgynevezett hét-transzmembrán receptorokon fejtik ki hatásukat, csakúgy, mint a jelenleg forgalomban levő gyógyszerek kb. 30-40 százaléka. Amennyiben a molekuláris struktúrából meg lehet jósolni, hogy egy adott molekula feltehetően milyen receptorhoz, fehérjéhez tud majd kötődni, az segíthet a jövőben újabb hatóanyagok megtalálásában, és a vegyületek hatásainak és mellékhatásainak a feltérképezésében. Ez felgyorsíthatja újabb gyógyszerek fejlesztését, és csökkentheti az azzal járó költségeket.

IRODALOM

Keller, A. – Gerkin, R. C. – Guan, Y. et al. (2017): Predicting Human Olfactory Perception from Chemical Features of Odor Molecules. *Science*, 24, 355, 6327, 820–826. DOI: 10.1126/science.aal2014, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5455768/

Keller, A. – Vosshall, L. B. (2016): Olfactory Perception of Chemically Diverse Molecules. *BMC Neuroscience*. 8, 17, 1, 55. DOI: 10.1186/s12868-016-0287-2, https://bmcneurosci.biomedcentral.com/articles/10.1186/s12868-016-0287-2

Pedregosa et al. (2011): Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Rogers, D. – Hahn, M. (2010): Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50, 742–754. DOI: 10.1021/ci100050t, https://www.researchgate.net/publication/43350565_Extended-Connectivity_Fingerprints

URL1: Open-source cheminformatics http://www.rdkit.org