

수리생물학: 우표 더미 속의 규칙을 찾는 법

최 지 범

고등과학원 계산과학부 위촉연구원

“모든 과학은 물리학이거나 우표 수집이다(All science is either physics or stamp collecting).”

—어니스트 러더퍼드(Ernest Rutherford)의 발언으로 종종 인용됨

많은 대학교의 수학과가 자연과학대학 소속이지만 엄밀한 의미에서 수학은 자연과학이 아니다. 공리(Axiom)와 논리 규칙을 활용해 새로운 정리(Theorem)를 찾아간다는 점에서 자연 현상의 관찰 결과에 의존하지 않기 때문이다. 따라서 고대 그리스의 피타고라스 정리를 비롯해 (증명에 오류가 없는) 수학 정리는 앞으로도 영원히 참이다. 반면 물리학, 생명과학과 같은 경험과학은 새로운 관찰 결과에 따라 기존의 법칙이 수정되거나 폐기될 수 있다. DNA에 대한 후생유전학 연구는 획득 형질이 유전되지 않는다는 오랜 믿음에 변화를 가져왔으며, 특정 진동수 이상의 빛만이 금속에서 전자를 방출시키는 광전효과는 파동이라 생각된 빛에 입자성이 있다는 것을 보였다. 토머스 쿤(Thomas Kuhn)에 따르면 자연과학은 이와 같은 패러다임 변화를 통해 발전한다.

자연과학의 한 축을 담당하는 물리학은 관찰을 기반으로 자연을 지배하는 법칙을 알아내는 학문이다. 수학은 관찰 결과와 독립적인 추상적 논리 체계이지만, 갈릴레오 갈릴레이(Galileo Galilei)에 따르면 ‘자연의 책’은 수학의 언어로 적혀있다. 자연법칙을 수학의 언어로 기술할 수 있기 때문이다. 아이작 뉴턴(Isaac Newton)은 달의 움직임을 보고 수학적 기법을 통해 만유인력의 법칙을 발견했다. m_1, m_2 의 질량으로 d 만큼 떨어진 두 물체 사이의 인력이 $G \frac{m_1 m_2}{d^2}$ 라는 이 법칙은 지구와 달, 인공위성과 행성의 움직임을 간결하게 표현할 뿐 아니라 앞으로 어떻게 움직일지도 정확하게 예측한다. 다시 말해 물리학자는 자연으로부터 영감을 얻고, 논리와 수식을 통해 그 속에 숨은 법칙을 발견하며, 이를 통해 자연을 예측하고 설명하는 이들이다. 양자역학 연구자가 웬만한 수학자만큼 수학을 잘하는 것도 이상한 일이 아니다.

반면 생명과학은 역사적으로 현상의 기록에 더 가까웠다. 한 지역에 어떤 종이 얼마나 서식하고 무슨 특징을 가지는지 조사하는 것은 육체적으로 고되면서도 가치 있는 일이지만, 그 안에서 일반적인 법칙을 찾기란 쉽지 않다. 우표 수집은 규칙을 찾는 것이 아닌 단순히 데이터를 모으는 것에 대한 비유 또는 풍자라 할 수 있다. 현대 생명과학의 주류인 분자 생물학 역시 관찰에 방점이 찍혀있다. 세포 내 분자적 상호작용을 분석하는 정교한 실험을 통해 많은 데이터와 유용한 지식이 축적될 순 있어도 만유인력의 법칙 같은 일반적 규칙을 찾기는 어렵다.

무엇보다도 대다수의 생명현상은 여러 요소가 영향을 미치기에 복잡하다. 암의 증식을 수학적으로 분석한다고 해보자. 이미 많은 연구자들이 암세포를 배양 접시에서 기르며 이들의 성장을 기록한 데이터를 갖고 있다. 이 데이터를 참고해 Gompertz 모델, Mendelsohn 모델, Bertalanffy 모델, Logistic 모델 등 암의 성장을 설명하는 여러 수학 모델이 제시되었다[1]. 이 식들은 3개 이내의 계수(Coefficient)를 통해 시간에 따른 암의 성장을 설명한다. 예컨대 Logistic 모델은 다음과 같은 형태이다(그림 1a).

$$\frac{dV[t]}{dt} = rV[t] \left(1 - \frac{V[t]}{K}\right).$$

$V[t]$ 는 시간 t 에서 암의 크기를 나타낸다. $dV[t]/dt$ 는 시간에 따라 암의 크기가 변하는 정도이며, 이 값이 양수이면 암은 성장하고, 음수이면 줄어든다. 암이 작다면($V[t] \approx 0$), 위 식은 $dV[t]/dt = rV[t]$ 와 비슷한 형태이다. 이 경우, 암의 성장 속도가 자신의 부피에 비례하는 기하급수적인 성장을 보인다. 크기가 증가함에 따라 $1 - V[t]/K$ 는 점차 0에 가까워져 암의 성장도 둔화되고 결국 작은 크기에서 시작한 암은 K 보다 커질 수 없다. 실험실에서 얻은 데이터 중 초기 성장 속도를 통해 r 값을, 최종 크기를 통해 K 값을 유추할 수 있기에 이 모델은 간단하면서도 실용적이라 할 수 있다(그림 1b-d).

그렇지만 실제 우리 몸에서 암이 자라는 과정은 실험실의 배양 접시보다 훨씬 더 복잡하다(그림 1e). 폐암, 췌장암, 뇌암 등 암의 유형에 따라 세포의 특성이 다를 뿐만 아니라, 동일한 폐암이라 하더라도 환자 간에 뚜렷한 차이를 보일 수 있다. 한 덩어리의 암 안에도 다양한 암세포가 군락을 이루고, 시간이 지나면서 암세포도 변화한다. 암세포를 발견한 우리 몸의 면역 시스템도 그에 맞춰 대응하고 새로운 전략을 수립한다. 예컨대 특정 면역 세포가 암세포를 제거하는 데 성공하면, 우리 몸은 그와 동일한 면역 세포를 추가적으로 투입해 암을 억제한다. 반대로 암세포의 입장에서는 면역 세포의 공격으로부터 살아남은 이들이 번성한다. 생태계의 적자생존을 연상시키는 면역 체계와 암세포의 상호작용으로 인해 암이 완전히 제거되거나, 특정 크기에 머물러 있거나, 증식을 거듭해 환자를 죽게 만들 수 있다. 우연적인 요소와 더불어 워낙 복잡한 요소들이 개입되었기에 이를 수학적으로 기술하는 것 역시 간단치 않다. 환자의 나이, 항암제 투여 여부, 암세포의 종류 및 다양성, 각 면역 세포의 활성화 등을 고려한 수학 모델을 굳이 만든다면 그 형태도 복잡하고 계수가 10개 이상은 필요할 것이다(그림 1e). Logistic 모델에서 최종적인 암의 크기를 통해 K 값을 유추하는 것과 달리 면역 세포의 구성 등에 대한 정보는 측정하기도 쉽지 않다.

컴퓨터를 통해 모델 결과를 구할 때도 복잡한 모델은 계산 부담이 훨씬 크다. 예컨대, 다음과 같이 5개의 인자를 갖는 모델이 있다고 하자.

$$y = f[x_1, x_2, x_3, x_4, x_5].$$

각 x_i 에 대해 0.1부터 1.0까지의 값을 0.1 단위로 대입한 후 y 를 구한다면, 총 10^5 번의 계산을 수행해야 한다. 인자가 10개라면 계산량은 10^{10} 으로 늘어난다. 각 인자가 하나의 차원을 만들기 때문인데, 이처럼 차원이 많아질수록 계산량이 기하급수적으로 증가하는 현상을 차원의 저주(Curse of dimensionality)라 부른다.

그렇다고 간단한 모델이 항상 유리한 것은 아니다. 암의 성장에 시간과 영양분, 산성도가 중요한 영향을 미친다고 하자. 시간과 영양분만 고려해 수리 모델을 만들었다면, 그 모델은 다양한 산성 환경에서 자라는 암을 제대로 설명하지 못한다. 산성도에 주목하지 않고 그 모델을 사용하는 실험 연구자는 왜 모델이 자신의 데이터와 맞지 않는지 혼란에 빠질 것이다. 장난감 모델(Toy model)이라고도 불리는 극단적으로 단순한 모델은 구(Sphere)와 소(Cow)를 동일시하는 것과 같다는 우스갯소리도 있다. 소 역시 하나의 연

결된 덩어리이니, 구와 다를 바 없다는 것이다.

다시 말해, 간단한 모델은 분석하기 용이하지만 복잡한 현실을 반영하기 어렵다. 반면 현실성을 반영한 모델은 때로는 정확할 수 있지만, 복잡하고 필요한 계수를 유추하기 까다롭다. 또한 복잡한 모델에서 특정한 패턴이 발견되었을 때, 어떤 인자가 그 패턴을 일으켰는지 파악하기 어렵다.

셀 수 없이 많은 요소가 영향을 미치는 생명현상을 수학의 언어로 분석하려면 단순함과 복잡함 사이에서 중용을 추구해야 한다. 중대한 요소를 무시해서도 안 되며, 불필요한 디테일에 얽매어서도 안 된다. 성향에 따라 균형 잡힌 복잡성에 대한 견해도 다르기에, 같은 현상에 대해서도 하나의 변수 혹은 10개 이상의 변수로 분석한 모델이 존재한다. 게다가 생명과학 모델을 만들기 위한 데이터가 충분치 않거나 오차가 큰 경우도 많다. 이런 이유 때문인지 수학이 자연과학 내의 영역을 확장하는 와중에도 생명과학에 수학을 적용하려는 시도는 많지 않았다.

생명과학자들이 성향상 수학을 별로 좋아하지 않는다는 점도 영향을 미쳤을 것이다. 팀 포셋(Tim Fawcett)과 앤드루 히긴슨(Andrew Higginson)이 2012년 미국 국립과학원회보(Proceedings of the National Academy of Sciences)에 발표한 논문[2]은 이런 성향을 구체적으로 보여준다. 이들의 조사 결과에 따르면 생태 및 진화 분야 논문에 더 많은 수식이 등장할수록 그 논문이 비이론(Nontheoretical) 논문에서 인용될 가능성이 유의미하게 감소한다. 반면 수식이 논문의 부록(Appendix) 형태로 등장하면 이런 경향이 발견되지 않았다. 이론 연구자와 실험 연구자의 소통은 과학 발전에 필수적이기에 인용 데이터를 분석한 저자들은 복잡한 수식을 부록으로 옮기는 것이 효과적일 수 있다고 제안했다.

이러한 어려움에도 불구하고 생명현상에 대한 수학적 분석은 점차 활기를 띠고 있으며, 인류에게 새로운 통찰을 제공했다. 동물의 뇌에서 수면과 관련된 특정 물질이 생산되면, 이 물질이 같은 물질의 추가적인 생산을 막는 줄임 되먹임(Negative feedback)이 일어난다. 그 결과, 해당 물질의 양도 줄어들고, 억제 요인 역시 사라지기에 생산량은 다시 증가한다. 이 과정이 되풀이되며 뇌는 대략 하루마다 반복되는 일주기 리듬(Circadian rhythm)을 만들어낸다. 물리학자는 진자의 운동과 같은 주기성을 설명하기 위해 미분방정식을 활용해 왔다. 미분방정식은 현재 상태가 앞으로의 변화에 어떤 영향을 주는지 수학적으로 설명하는 기법이며, 앞서 살펴본 Logistic 모델 역시 미분방정식의 한 예이다. 각 뇌세포가 어떤 물질을 만들고, 그 물질이 무슨 작용을 하는지 알고 있다면, 이를 수학적으로 해석해 미분방정식에 녹여낼 수 있다. 이렇게 만든 수리 모델은 기존 데이터를 설명할 뿐 아니라 뇌가 주기성을 만드는 원리를 제시하며, 발견되지 않은 현상을 예측할 수 있다. 카이스트 수리과학과의 김재경 교수 연구팀은 이 현상에 대한 미분방정식을 개발하고 실험을 수행해, 어떻게 뇌가 안정적이면서도 탄력적으로 일주기 리듬을 조절하는지 규명했다[3]. 이 연구를 활용해 수면장애에 대한 실질적인 치료법을 고안할 수 있을 것이다.

인간의 세포핵에는 30억 쌍의 DNA 염기서열이 있고, 가까운 친족일수록 염기서열 구성도 비슷하다. 방대한 양의 데이터지만, 주성분 분석(Principal component analysis)이라는 수학적 기법을 사용하면 단 2개의 축으로 개인의 염기서열 구성을 표현할 수 있다. 2개의 축으로 간소화된 평면에 다양한 인간의 DNA 서열 정보를 표시하면, 이들이 유전적으로 얼마나 가까운지 파악할 수 있다. 집단유전학자들은 고대 인류의 DNA를 이런 방식으로 분석해 인류의 진화와 이동의 역사를 규명했다[4]. 예컨대 물리적으로 먼 곳에 위치한 두 선사시대 부족이 서로 비슷한 유전 구성을 보인다면, 원래 부족의 일부가 먼 거리를 이동해

새로운 곳에 정착했다는 것을 의미한다.

방대한 DNA 서열 정보와 인간의 일주기 리듬은 창고를 가득 채운 우표라 할 수 있다. 규칙이 보이지 않을 것 같은 그런 우표 더미 속에서 정교한 수리 모델은 본질을 관통하는 원리를 찾게 해준다. 이를 위해서는 서로 다른 세계에 사는 수학자와 생명과학자들이 소통하고 교류할 필요가 있다. 다른 파장의 전자기파를 관측하는 망원경으로 밤하늘을 보면 전혀 다른 우주가 펼쳐지듯, 수학과 생명과학이 잉태한 수리생물학의 시선으로 생명을 바라보면 새로운 패러다임이 우리에게 다가온다.

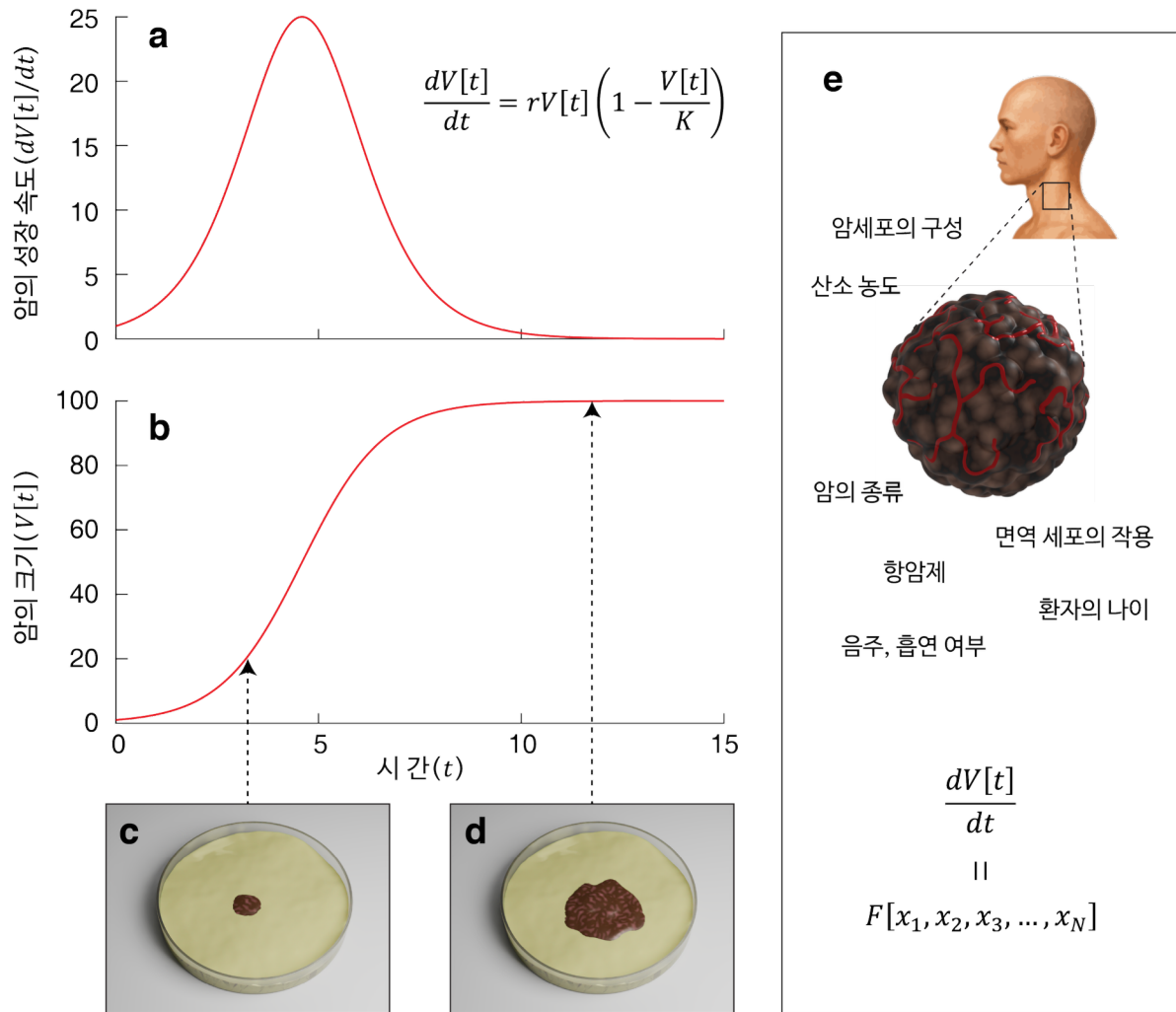


그림 1. 암의 성장에 대한 수리적 모델

(a) Logistic 모델에 따르면 암은 점차 빠르게 성장하다 이후 그 속도가 둔화한다. 이 그래프는 $r = 1$, $K = 100$ 인 상황을 보여준다.

(b) 모델에서 구한 성장 속도를 통해 시간에 따른 암의 크기를 예측할 수 있다.

(c, d) 실제 암의 성장 데이터를 모델 결과와 비교하면 모델의 타당성을 검증할 수 있다.

(e) 간결한 수학 모델과 달리 실제 인간 몸에서 자라는 암은 여러 요소에 영향을 받으며, 그 데이터를 얻기도 쉽지 않다.

참고 문헌

1. Gerlee, P. (2013). The model muddle: in search of tumor growth laws. *Cancer Research* 73, 2407–2411.
2. Fawcett, T.W., and Higginson, A.D. (2012). Heavy use of equations impedes communication among biologists. *Proceedings of the National Academy of Sciences* 109, 11735–11739.
3. Jeong, E.M., Kwon, M., Cho, E., Lee, S.H., Kim, H., Kim, E.Y., and Kim, J.K. (2022). Systematic modeling-driven experiments identify distinct molecular clockworks underlying hierarchically organized pacemaker neurons. *Proceedings of the National Academy of Sciences* 119, e2113403119.
4. Jeong, C., Wang, K., Wilkin, S., Taylor, W.T.T., Miller, B.K., Bemmman, J.H., Stahl, R., Chiovelli, C., Knolle, F., and Ulziibayar, S. (2020). A dynamic 6,000-year genetic history of Eurasia’s Eastern Steppe. *Cell* 183, 890–904.