A Galaxy of Cows, a Constellation of Chairs

Mike Arnautov

A few years ago, our treasurer Eileen Walker asked me a really difficult question: what is a cow? One does not need to be a philosopher to spot that this is a tricky one. Short of finding a cow and pointing at it, how does one define a cow?. So I resorted to quoting my favourite philosopher politician (which is the best we can do these days, in the absence of philosopher kings). During the famous episode during his presidency, Bill Clinton parried an equally difficult question ('Well, is it or isn't it?', if I recall correctly) with a profoundly philosophical observation: 'That depends on what the meaning of 'is' is.' I don't think he ever got the credit he deserved for injecting such a deep thought into the political process.

Seeing that I wouldn't bite, Eileen switched her tactic: well then, what makes a cow a cow? I felt a bit more comfortable with that, but still unwilling to commit myself. After all, this question used to have at least one pretty standard answer: it is the phylogenetic tree – the Darwinian 'tree of descent' that makes an organism what it is. If only it really were so simple! Now that we understand just how promiscuous DNA can be[1], the famous tree turns out to be more of a web. And as if that weren't enough, we humans are busily making things worse. If a cow has a bison for its surrogate mother, is it still a cow? Or if Craig Venter succeeds in his ambition to assemble a bacterium from scratch what would be its tree of descent – humans?

So I dodged again. But Eileen would not give up: tell me at least, how do you think we know that a cow is a cow? Ah! Now that one I think I can answer. We know that a cow is a cow the same way as we know that a tree belongs to a particular copse (wood, forest... whatever). I was happy with that answer, but Eileen was not, so I attempted to explain. That explanation is the basis of my talk today.

However, before I proceed, I should face up to one likely objection. Some of my listeners may complain that in considering the question 'how do we know that a cow is a cow?' I am concerned with epistemology, whereas the theme today is supposed to be metaphysics, which is often taken to mean ontology. It is, of course, a standard injunction in philosophy to keep one's ontology and epistemology separate, but I feel that it would be a mistake to interpret this worthy slogan too simplistically. The truth of the matter is that one cannot talk about ontology without taking epistemology into the account ('you say that so-and-so is/isn't – how do you know that to be the case?') or vice versa ('you say you are considering how we can know something – what sort of something do you have in mind?'). This is not to say that ontology could be reduced to knowability. It's just that the two cannot be cleanly separated.. Hence I understand the advice not to mix the two as meaning not to mix them at the same level of importance. If ontology is our theme, let epistemology be its servant.

I should also make it clear that while the question was asked specifically about cows, I interpret it generally as a query about classifying living beings (not just cows) into related groups, such as species. How do we classify plants and animals into species? A traditional answer, one I was taught at school, is that the key is in interbreeding. A cow and a bull can breed, so they belong to the same species. Two bulls can breed with the same cow, so they too belong to the same species etc.

While this generally works well enough, this definition has its problems.

---

1 It is now accepted that at least some of our genes (and some researchers say a substantial fraction of our genes) do not originate in the human phylogenetic tree. For some discussion of this phenomenon, known as Horizontal Gene Transmission see, e.g., http://en.wikipedia.org/wiki/Horizontal_gene_transfer

- Surely, a cow is a cow even before it is sexually mature.
- How did the first cow reproduce?
- Not all living beings reproduce sexually.
- There are inconvenient curiosities such as 'ring species'[2].
- Interspecies cross-breeding is rare but not unheard of .

In any case, we mostly need the answer fairly quickly, without having an opportunity to enquire into an organism's breeding habit. Some other approach is required when trying to answer questions such as 'is this berry safe to eat?' or 'is that a tiger there, or a zebra?'. Clearly, that's a sort of question humans have always been able to answer – we would not be here otherwise. Equally clearly, our identification methods, quick and dirty as they may be, depend on noting some easily observable features of the organism in question, and using those features as a basis for classification.

Unsurprisingly, that's in essence the method adopted by science. Zoologists and botanists have been classifying fauna and flora for quite some time now. How do they know what species to assign to a given individual? The art of such classification (and make no mistake, it is an art) depends on selecting some small or even minimal set of defining features, which in combination are taken to be uniquely characteristic of a species.[3]

A philosopher may well object that in doing so, scientists are just as likely to impose classifications which have little if any relationship to 'natural kinds'. However, after some considerable arguments botany and zoology settled on agreed classification criteria and the system worked well enough in the sense of producing practical results. It was therefore generally felt that, whatever philosophers might think, such classifications were good enough in carving the kingdom of living beings at its joints.

The discovery and the eventual decoding of the genetic aspect of life strongly suggested that the genetic inheritance of an organism (its genome) dictated the bodily expression of that organism (its phenome), subject to interference from external factors. It was therefore logical to expect that species delineation obtained by older morphological methods should match whatever delineation can be derived genetically, i.e. by studying structural similarities of genomes.

This work is still ongoing, but to everybody's relief it turns out that, despite a few well publicised botanical upsets, traditional species classifications stands up well to these checks. Admittedly the relationship between genome and the somatic characteristics of an organism (its phenome) turned out to be rather more complex then expected. We now know that epigenetic (i.e. non-genetic) inheritance also plays a role – so much so that it may make sense to think of genome and its corresponding phenome as being a result of co-evolution, rather than there being just genetic evolution and the corresponding succession of phenome expressions. We also now know that there can be a significant amount of genome variation across the body of an individual organism[4].

---

2 To quote Wikipedia: In biology, a ring species is a connected series of neighbouring populations, each of which can interbreed with closely sited related populations, but for which there exist at least two 'end' populations in the series, which are too distantly related to interbreed, though there is a potential gene flow between each 'linked' species. Such non-breeding, though genetically connected, 'end' populations may co-exist in the same region thus closing a 'ring'. (http://en.wikipedia.org/wiki/Ring_species)

3 The eventual agreement on preferred taxonomies was arrived at after some prolonged and heated debates. E.g. in classifying plants: what matters more: overall body plan, the foliage, flowering morphology or fruiting morphology? Thus the old question: should one differentiate between trees and herbs? (The modern answer is 'no'.)

4 E.g. it has been recently reported that there is often a greater genome variability between different parts of cottonwood trees than between corresponding parts of different trees. (http://www.nature.com/news/tree-s-leaves-genetically-different-from-its-roots-1.11156)

Fortunately, none of that makes a real difference to the answer I gave Eileen, because it does not affect the underlying methodology widely used in structural comparisons of genomes. The method in question is called 'cluster analysis' and it takes us back to knowing that a tree belongs to a particular group of trees. At least that's my personal visualisation of the method. In trying to find the best way of explaining it to a non-technical audience, I eventually decided that interstellar exploration is probably a more appropriate analogy. Doubly so in that it provides a natural link to the galaxy in the title of this talk. Don't worry for the moment where cows fit into that – I'll return to them later.

The cluster analysis technique is simple enough in its essence. It is this essence I shall try to explain – not the actual ways (highly mathematical and/or algorithmic) in which the technique is used in practice. We can all do clustering quite unconsciously when looking at visual data. If there is a group of dots in the top right corner of a piece of paper and another in the bottom left, we simply see them as two distinct clusters of points. But how does one do the same trick with data, such as genomes, which cannot be represented in such visual manner? The trick is to use a visually simple case, and work out a satisfactory non-visual approach, which can then be used on non-visualisable data..

So... A galaxy of stars. To avoid getting bogged down in details (without however, sacrificing rigour), let's envisage our Milky Way galaxy as a very clean and uniform multi-arm spiral with two blob-like satellite galaxies (i.e. the two Magellanic clouds) off to one side. Now, suppose you have a miraculous space ship which can make any number of instantaneous interstellar jumps provided (a) the jump is no longer than some particular number of light years (call this number R for 'Reach') and (b) the end point of a jump must be in a close proximity to a star – say within a few light hours at most. Given a specific value of R, how many stars could you visit? If R is less than 4 light years, you are stuck in the Solar system. But inventors of this miraculous vehicle are busily improving its performance and R keeps growing as time goes by. Once you can hop to Alpha Centauri which is 4.2 light years away, you can also visit any star within 4.2 light years of it and then any star within 4.2 light years of that and so on. As R grows, the number of stars within reach of your fantastic starship will grow very rapidly. Even before it can jump across the gaps separating spiral arms of the galaxy, other galaxy arm will be reachable by travelling along our arm towards the galaxy core and then out again along another arm.. In fact a relatively modest value of R, something like a few tens of light years would allow you to visit every star within the galaxy, despite the galaxy's size – about 100,000 light years in diameter – being vastly greater than the ship's reach .

It is interesting to plot the number of stars N one can reach, against the value of R. The trend will be steeply upwards, with N growing very rapidly from 1 to the full 300 billion or so – the total number of stars in the Milky Way. What happens once every star in the galaxy is within your reach? R keeps growing, yet there are no more new stars to be visited. That's because the Milky Way Galaxy is a well delineated cluster of stars. The value of R at which N stopped growing, i.e. the minimal jump size which allows you to visit every star is the *binding distance* of the galaxy

But let R grow further still and suddenly the value of N leaps upwards again. What has happened? Your starship can now jump across the narrowest gap between our galaxy and one of it satellite mini-galaxies . And, since the two satellite galaxies are closer to each other than they are to the Milky Way galaxy, both become immediately accessible in full. The value of R at which this happens is the binding distance for the cluster of the Milky Way and the two Magellanic clouds. Then there will be a long time of R growing without any noticeable effects, until one gets to jump all the way to the Andromeda galaxy. And so on...

The step-like nature of our plot of N against R, with its sudden jumps, often coming after a period of no change, is characteristic of clustering of stars into large-scale groups. It can also reveal a hierarchical aspect of clustering. If our journey started not from within the main galaxy but from within one of the two Magellanic clouds, we would see a slightly different pattern. Again, N would quickly grow to the complete count of stars in that cluster of stars. Then after a pause it would jump again, as the other Magellanic cloud comes into range, because the two of them are closer to each other than to the main galaxy itself. Then after another pause, there would be another large jump – the main galaxy is now reached too.

Thus not only can we tell that stars are clustered, but we can also work out a hierarchy in that clustering. Thee two Magellanic clouds can be grouped together by a smaller binding distance before bringing Milky Way itself into the group.

There are several interesting things about this way of looking at clusters via their binding distance:

- It is completely insensitive to the shape and the size of the cluster – a small group of stars can have the same binding distance as, for example, a very long linear strip of stars.[5]
- It signals the existence of separate clusters by the phenomenon of a sudden jump in the number of available destinations, often (but not necessarily) after a period of R growing without any effect.
- It depends on nothing more than a notion of distance between a pair of stars.

That has a significant consequence. Binding distance allows us to differentiate between clusters which are hard to differentiate otherwise. E.g. two clusters arranged as two long strips, with both strips much longer than the gap between them. Or two C-shaped strips with end inside each other. Or a circular strip with a separate cluster inside the circle, And so on...

Of course, it is possible to think of rather special configurations of stars relative to one's starting point, for which things do not work out so cleanly. However, this is only the case in the very simplified picture I have outlined here. In actual applications of the technique there is no one particular point treated as the origin. We have already done something similar in considering what structure is revealed by choosing a starting point within one of the Magellanic clouds instead of within the Milky Way.

This is all very well, but how does one apply all of this to cows, or to be more specific, to genomes? As the above explanation shows, all I really need is to establish that it is possible to define something like a sensible notion of 'distance' between any two genomes. The simplest way of doing this is to define such distance as the minimal number of 'atomic' modifications that would convert one genome into the other. This is in principle no different from what a spelling checker does when it presents you with a ranked list of possible word alternatives to a string of characters it fails to recognise as a word. Admittedly genomes are significantly longer, but there are surprisingly efficient algorithms which allow computers to calculate such genomic distance quite quickly. The whole discipline of bioinformatics is based on this.

Once we have such a notion of genomic distance[6], we can adapt our fantastic spaceship to 'fly' (in

5 Nevertheless, cluster analysis can be also deployed on cross-sections of data, yielding information on structural characteristics of a cluster. In this application of the technique such additional information comes from relationships between cluster structures revealed in various cross-sections.

6 Mathematicians may like to know whether such genomic distance is a mathematically well behaved metric. As can

some imaginary abstract multi-dimensional space, which we do not need to worry about) between genomes instead of between stars. Starting from a particular genome, for a particular value R of the ship's reach, how many other genomes are accessible?

It should come as no surprise that playing the spaceship game in this genome space shows patterns of the same kind as we see when exploring stars. In fact, it would be a major surprise *not* to find clustering identifiable by this method, once a sensible distance measure is defined. A lack of clustering would indicate a remarkable degree of regularity in spacing between objects being considered, and we have no reason to expect such regularity in the case of genomes.

Genomes are, as already noted, not the complete story. But this does not matter. It is perfectly possible to add traditional morphology as an adjunct to the genome space and define some suitable combined distance measure. A lot of current argument in genetics is about the precise details of measures to be used, but the overall outcome is clear enough. In most cases, clusters revealed by such analysis pretty well match the ways in which we divide living being into groups such as, e.g., species.[7]

By now you can probably anticipate my reason for saying that cows and other species are like galaxies. They form objectively identifiable clusters characterised by their specific binding distances and delineated by binding distance jumps from clusters of other species. In addition, results of cluster analysis conform to familiar hierarchies: on the one hand within the 'galaxy' of cows there are lesser clumps of different cow breeds, while on the other, we also find neighbouring 'galaxies' of, e.g., bisons, which together with cows form a 'meta-galaxy' of bovines.

Hence my conclusion is that in a well defined 'real' sense cows do exist and we have objective methods to decide whether a given organism is or is not a cow.[8] The same applies generally to all biological organisms. What is more, this approach cashes out in allowing us to tackle some of the oddities I have mentioned before.

- Yes, a sexually immature cow is a cow, and so is a cow embryo.
- Ring species are indeed species despite their odd breeding characteristics.
- A cow whose surrogate mother is a bison is definitely a cow.
- An E.Coli bacterium assembled from scratch in a lab is still E.Coli.
- There never was a 'first cow'; as Darwin knew without putting it like this, speciation happens by a cluster splitting into two or more, as some connecting members (genomes) disappear.

Having got this far with biology, it seem natural to ask whether it would be possible to use the same method for delineating inanimate objects. Specifically, can this approach assist with the philosophically most troublesome inanimate case of artefacts? At first glance it would seem that if we adopt the same approach looking at all measurable and/or classifiable properties of an object, then some sensible notion of distance might be derived, making cluster analysis possible. Unfortunately, there are at least two separate problems with this approach.

---

be readily seen, it satisfies the four standard criteria: non-negativity, identity of indiscernibles, symmetry and subadditivity. Hence the answer is yes.

7 It should be noted that the specific level of hierarchical clustering that we associate with a species, as opposed to a breed or a family, is an arbitrary one. This does not, however negate the fact that a species, on whatever level we decide to apply the term, is an identifiable cluster in the genomic space.

8 Note that there is a degree of feedback here. We take an organism, find a cluster of organisms to which it belongs, label that cluster 'cows' and hence call the organism in question a cow. In other words being a cow is a collective, not an individual characteristic.

Our recent technology does not help. I have a wireless router on my desk. But exactly the same physical box can be a bridge rather than a router, or a repeater, or an access point. It all depends on the software it is running. Strictly speaking, that software is a collection of binary bits, and as such it is a part of the physical description of the box, but there are so many possible different ways to construct such software that there any attempt to work out on that basis which of the possible labels to apply is really a hopeless enterprise. Manifestly such electronic boxes are classified not by what they are but by what they do, which makes the approach we used for living beings at best problematic.

We should not imagine that these difficulties are restricted to new-fangled electronics. I am reminded of the age-old question Zen masters ask novice monks: wherein lies the cartness of a cart? The correct answer is: the cartness of a cart is in its use. I suggest that this principle largely applies to many (most?, all?) artefacts. Take a humble chair – an object much beloved by philosophers. What exactly **is** a chair? Well, it is not a stool, so it must have a back. It generally has four legs, but could have other leg numbers (or even no legs at all – a stone cube with a back is also a chair). What makes a chair a chair? In the end it is its use, or potential use. Except... if a juggler juggles chairs, are they still chairs? Is a chair used as an improvised table still a chair? Perhaps it is the intended use that matters. But then if humans die out, is a chair still a chair? Or is it just a 'ritual object' to some Martian xenologist?

In any case problems do not stop there. Is an easy-chair a chair? It is not in Russian or Czech. Is a stool a chair? Not in English, but it is a 'small chair' in Czech (as is what an English speaker would call a small chair).

It is a well known linguistic fact that different languages delineate the cluster of objects possessing 'chairness' in different ways. While there is a central core on which they agree, what else is included or excluded differs from one language to another. There is no exact correspondence between 'chair' in English, 'stuhl' in German, 'chaise' in French or 'židle' in Czech.

This situation is strongly reminiscent of stellar constellations. In a sense we do group stars by use – by our projection into the starry sky of patterns and images. We may all agree that the Little Dipper is a distinct group of bright stars, but as to what else is to be included in Ursa Minor, that's very much an open question, dependent on the imaginary use we make of stellar patterns. And furthermore, those patterns themselves depend on the cosmic viewpoint from which we view them..

Hence my claim that cows (and biological species in general) are like galaxies, in that they do form objectively observable clusters. Chairs, on the other hand, and artefacts in general are more like constellations in that their extensions are less clearly defined and, to some extent at least, depend on one's point of view.

- o O o -